Multiple Sound Source Counting and Localization Based on TF-Wise Spatial Spectrum Clustering

Bing Yang, Hong Liu, Cheng Pang, and Xiaofei Li

***Abstract*—This paper addresses the problem of multiple sound source counting and localization in adverse acoustic environments, using microphone array recordings. The proposed time-frequency (TF) wise spatial spectrum clustering based method contains two stages. First, given the received sensor signals, the spatial correlation matrix is computed and denoised in the TF domain. The TF-wise spatial spectrum is estimated based on the signal subspace information, and further enhanced by an exponential transform, which can increase the reliability of the source presence possibility reflected by spatial spectrum. Second, to jointly count and localize sound sources, the enhanced TF-wise spatial spectra are divided into several clusters with each cluster corresponding to one source. Sources are successively detected by searching the significant peaks of the remaining global spatial spectrum, which is formed using unassigned spatial spectra. After each new source detection, spatial spectra are reassigned to detected sources according to the dominance association between them. The interaction between sources is reduced by iteratively performing new source detection and spatial spectrum assignment. Experiments on both simulated data and real-world data demonstrate the superiority of the proposed method for multiple sound source counting and localization in the environment with different levels of noise and reverberation.**

***Index Terms*—Source counting, multiple sound source localization, TF-wise spatial spectrum clustering, signal subspace.**

I. Introduction

MULTIPLE sound source localization using microphone arrays is crucial for numerous acoustic signal processing tasks, such as speech dereverberation, noise reduction and blind source separation. In recent decades, a wide range of approaches have been developed for source localization, which can be roughly classified into two categories: spatial spectrum [1]–[3] and time-frequency (TF) processing [4], [5]. Despite the great progress with these methods, the localization performance is still greatly affected by unknown source number, adverse acoustic conditions, etc. [6]–[8].

Manuscript received August 20, 2018; revised March 6, 2019 and April 30, 2019; accepted April 30, 2019. Date of publication May 10, 2019; date of current version May 20, 2019. This work was supported by the National Natural Science Foundation of China under Grant 61673030 and Grant U1613209. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Roland Badeau. *(Corresponding author: Hong Liu.)*

B. Yang, H. Liu, and C. Pang are with the Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, Beijing 100871, China (e-mail: bingyang@sz.pku.edu.cn; hongliu@pku.edu.cn; chengpang@sz.pku.edu.cn).

X. Li is with the INRIA Grenoble Rhône-Alpes, Montbonnot-Saint-Martin 38330, France (e-mail: xiaofei.li@inria.fr).

Digital Object Identifier 10.1109/TASLP.2019.2915785

The TF processing methods exhibit the ability to jointly count and localize multiple sound sources. They generally assume that at most one source is dominant over the others in the TF domain due to the sparse property of speech signals. This assumption, namely the W-disjoint orthogonal (WDO) assumption [9], can simplify multiple source localization on broadband to single source localization in individual TF bins. When combining TF processing with spatial spectrum for effective multiple sound source localization, two issues need to be considered in general: 1) How to build reliable TF-wise spatial spectrum to indicate the source presence by its significant peak. 2) How to accurately localize multiple sound sources according to TF-wise spatial spectra, especially when the number of sources is unknown.

For individual TF bins where one source is dominant, various methods have been exploited to build spatial spectrum as a function of source location, such as steered response power (SRP) [1] and multiple signal classification (MUSIC) [2]. Among them, the signal subspace information is investigated intensively, and the eigenvectors of spatial correlation matrix (SCM) are demonstrated to exhibit different properties. Generally, when SCM is not considerably contaminated by noise, the principal eigenvector encodes the steering vector of one source, and spans the signal subspace of this source [10], [11]. The remaining eigenvectors span the noise subspace and exhibit orthogonality to the steering vector of the source [2], [12]–[14]. Different from the structure of SCM, parameterized SCM is designed as a function of candidate locations [15], [16]. The cue for localization is that when parameterized SCM is steered toward the true source location, the largest eigenvalue is maximized while all the other eigenvalues are minimized [16]. An alternative cue is that the determinant of parameterized SCM reaches its minimum at the true source location [17], [18]. Though subspace methods can achieve high-resolution localization for the fine division of candidate location space, most of them are sensitive to acoustic interference, since SCM could be distorted by interference.

Robustness to acoustic interference is necessary for reliable multiple sound source localization. Many processes are designed to improve the robustness of localization to ambient noise, room reverberation, interaction between sources, etc. Ito *et al.* utilized the geometrical information of symmetric microphone arrays [19] to improve the MUSIC performance in environments with diffuse noise [20]. Huang *et al.* exploited the precedence effect [21] to robustly estimate direction of arrival (DOA) of each source in reverberant scenarios [22]. Li *et al.* used the segmental SCM subtraction method to obtain a noise-free SCM for single speaker relative transfer function identification [23]. Pang

et al. proposed a reverberation weighting method to separately suppress the early and late reverberation while preserving the localization cues [24]. Several processes dedicate to identifying single source dominated TF bins, such as coherence test [25], single source confidence measure [6] and direct path dominance test [26]. They usually serve as the preprocessing stage to TF processing methods, with the goal to discard the TF bins which are severely affected by multiple sources or reverberation. Motivated by these works, the signal subspace based spatial spectrum is improved in this work to have an adaptability under different adverse acoustic conditions.

In general, multiple source localization can be conducted in three ways given TF-wise spatial spectra [27]. 1) Histogram method [6], [11], [28]–[30]: Source location is estimated from each TF-wise spatial spectrum, and these estimations are collected to form a histogram which is used for localization. 2) Global method [26], [27], [31], [32]: TF-wise spatial spectra of all TF bins are united to obtain the global spatial spectrum from which locations of sources are estimated. 3) Source-associated method [10], [33]–[35]: For each source, associated TF-wise spatial spectra are identified to construct the single-source global spatial spectrum, and the source location is estimated accordingly.

To tackle the case where the number of sources is unknown, a common approach is to determine the number and locations of sources by detecting significant peaks of location histogram or global spatial spectrum [11]. The methods in [6], [34], [35] count and localize sources one by one in a similar way to the matching pursuit algorithm [36], which detect and remove contribution of each source alternately until a stop criterion is satisfied. In [6], each source is detected by maximizing the correlation between direction-centered smooth pulse and DOA histogram, and then the pulse-shaped histogram corresponding to the detected source is removed. In [34], [35], each source is detected by searching the highest peak of the global spatial spectrum, and then the TF bins associated with the detected source are identified and removed. Additionally, the iterative source counting and localization in [35] is followed by a DOA refining process which iteratively optimizes the TF bins used for localizing each source. Though favorable performance can be achieved by these methods, the performance is still affected by the interaction between sources especially when sources are close to each other.

In this paper, we propose a novel multiple sound source counting and localization method based on TF-wise spatial spectrum clustering. The proposed method has the following procedures and contributions.

First, the spatial correlation matrix (SCM) is estimated in the TF domain using the received sensor signals. To reduce the signal subspace distortion caused by diffuse noise, the SCM denoising method [37] which is originally designed for MUSIC is adopted. Considering the geometrical configuration of symmetric microphone array and TF sparsity of speech signals, the noise-free SCM is recovered by removing the diffuse noise from the noise-contaminated SCM. For each single source dominated TF bin, the similarity between the principal eigenvector of the noise-free SCM and the steering vector related to all candidate directions is calculated to estimate the TF-wise spatial spectrum. To make the spatial spectrum more reliable to reflect the source presence possibility, an exponential transform is proposed, which can suppress indistinctive peaks and preserve significant peaks simultaneously.

Second, to jointly count and localize multiple sound sources, a TF-wise spatial spectrum clustering algorithm is designed, which contains two nested iterative procedures with the internal iterative procedure serving as a step of the external iterative procedure. In the external iterative procedure, sources are successively detected based on the spatial sparsity of sources. After the detection of one source, the DOA information of all detected sources are utilized to initialize the internal iterative procedure where the spatial spectra dominated by different detected sources are respectively distinguished and assigned to them. Then the unassigned spatial spectra are employed for the next source detection. The clustering algorithm divides the spatial spectra into several clusters. The number and DOAs of sources are respectively determined by the number of clusters and the spatial spectra in each cluster. The dominance association between spatial spectra and sources, which is adjusted by the internal iterative procedure, helps to reduce the interaction between detected sources from a spatial spectrum perspective. In addition, the influence of already detected sources on the detection of other sources is reduced, since spatial spectra dominated by detected sources have been removed and only remaining spatial spectra are used for new source detection. Hence, multiple sound sources are reliably counted, and their DOAs are estimated by fusing the direction information associated with each source.

Overall, the proposed method is robust under different acoustic conditions, since SCM denoising and exponential transform are employed to increase the reliability of TF-wise spatial spectrum, and the spatial spectrum clustering algorithm is designed with the tolerance of interaction between sources. Furthermore, the proposed method can simultaneously estimate the number and DOAs of multiple sources, which is demonstrated to perform better than several other methods on both simulated and real-world data. Though we mainly focus on counting and localizing speech sources in this paper, the proposed method can also be applied to other audio sources whose signals are sparse in the TF domain, such as some types of musical sources.

The remainder of this paper is organized as follows. Section II formulates the signal model for multiple sound source counting and localization. Section III describes the TF-wise spatial spectrum estimation. Section IV details the TF-wise spatial spectrum clustering algorithm. Experiments and discussions with simulated and real-world data are presented in Section V, and conclusions are drawn in Section VI.

II. Signal Model

Consider K far-field speech sources observed by a uniform circular array of M microphones in a noisy scenario as shown in Fig. 1. The signal received by the mth microphone can be modeled as:

$$x_m(t) = \sum_{k=1}^{K} \alpha_{mk} s_k(t - \tau_{m,\theta_k}) + v_m(t), \tag{1}$$

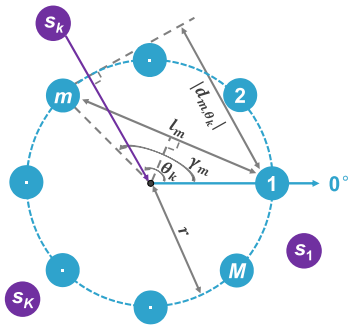


Fig. 1. Illustration of geometrical relationship, for a scenario where a uniform circular array with M microphones observes K far-field sound sources. The origin is defined at the center of the array.

where $m \in \{1, 2, \ldots, M\}$ is the microphone index, $k \in \{1, 2, \ldots, K\}$ is the source index, $s_k(t)$ denotes the signal emitted from the kth source, θ_k denotes the horizontal DOA of the kth source observed in an anti-clockwise manner with respect to 0°-ray, and $v_m(t)$ is the additive ambient noise received by the mth microphone which is assumed to be uncorrelated with the source signals. Here, α_{mk} and τ_{m,θ_k} represent the propagation attenuation factor and the time of arrival from the kth source to the mth microphone, respectively. By applying the short-time Fourier transform (STFT) to Eq. (1), the received signal can be modeled in the TF domain as:

$$X_m(n,f) = \sum_{k=1}^{K} \alpha_{mk} S_k(n,f) e^{-j\omega_f \tau_{m,\theta_k}} + V_m(n,f), \tag{2}$$

where n is the time frame index, f is the frequency band index, ω_f is the angular frequency of the fth frequency band, and $X_m(n,f)$, $S_k(n,f)$, $V_m(n,f)$ represent the STFT coefficients of $x_m(t)$, $s_k(t)$, $n_m(t)$, respectively. When expressed in a vector form, the signal model can be rewritten as:

$$\mathbf{x}(n,f) = \sum_{k=1}^{K} S_k(n,f)\mathbf{e}(f,\theta_k) + \mathbf{v}(n,f), \tag{3}$$

where

$$\begin{aligned}
\mathbf{x}(n,f) &= [X_1(n,f), X_2(n,f), \ldots, X_M(n,f)]^T, \\
\mathbf{v}(n,f) &= [V_1(n,f), V_2(n,f), \ldots, V_M(n,f)]^T, \\
\mathbf{e}(f,\theta_k) &= [\alpha_{1k} e^{-j\omega_f \tau_{1,\theta_k}}, \alpha_{2k} e^{-j\omega_f \tau_{2,\theta_k}}, \ldots, \\
&\qquad \alpha_{Mk} e^{-j\omega_f \tau_{M,\theta_k}}]^T.
\end{aligned}$$

Here, $\mathbf{e}(f,\theta_k)$ represents the steering vector related to the DOA θ_k, and $(\cdot)^T$ denotes the transpose operation.

The propagation attenuation factors α_{mk} for all sources and microphones are assumed to be identical, and denoted by α. Accordingly, taking the first microphone as the reference, the steering vector can be expressed as:

$$\begin{aligned}
\mathbf{e}(f,\theta_k) = \alpha e^{-j\omega_f \tau_{1,\theta_k}} \times [1, e^{-j\omega_f (\tau_{2,\theta_k} - \tau_{1,\theta_k})}, \ldots, \\
e^{-j\omega_f (\tau_{M,\theta_k} - \tau_{1,\theta_k})}]^T.
\end{aligned} \tag{4}$$

For far-field model where the propagation paths from one sound source to different microphones are regarded to be parallel, the relative time delay between signals captured by the mth microphone and the reference microphone for the kth source is computed as:

$$\tau_{m,\theta_k} - \tau_{1,\theta_k} = \frac{d_{m,\theta_k}}{c} = \frac{l_m \sin(\gamma_m/2 - \theta_k)}{c}. \tag{5}$$

Here, as depicted in Fig. 1, d_{m,θ_k} is the distance difference from the kth source to the mth microphone and the reference, l_m is the distance between the mth microphone and the reference, and γ_m represents the angle of the mth microphone with respect to 0°-ray. Considering the geometrical relationship between microphones, γ_m and l_m are calculated as:

$$\gamma_m = (m-1)\frac{2\pi}{M}, \quad l_m = 2r\cos\left(\frac{\gamma_m}{2} - \frac{\pi}{2}\right), \tag{6}$$

where r denotes the radius of the microphone array.

Generally, the number of sources is unknown in practice. Hence, given the sensor signals and the microphone array geometry, the task is to simultaneously estimate the number K and the DOAs of sound sources $\{\theta_1, \ldots, \theta_K\}$.

III. TF-Wise Spatial Spectrum Estimation

In this section, we estimate the spatial spectrum for the single source dominated TF bins in the presence of acoustic interferences. For clarity, we first briefly present the related theoretical basis. Then, the spatial correlation matrix (SCM) is calculated and denoised without affecting the signal subspace. Finally, the TF-wise spatial spectrum is estimated based on the signal subspace information and further enhanced by the exponential transform.

A. Preliminaries

With $\mathbf{x}(n,f)$, the SCM for each TF bin is given by:

$$\mathbf{R}_{\mathrm{xx}}(n,f) = E\left\{\mathbf{x}(n,f)\mathbf{x}^H(n,f)\right\} = \mathbf{R}_{\mathrm{cc}}(n,f) + \mathbf{R}_{\mathrm{vv}}(n,f), \tag{7}$$

where $E\{\cdot\}$ denotes expectation, $(\cdot)^H$ denotes the conjugate transpose operation, $\mathbf{R}_{\mathrm{vv}}(n,f)$ represents the noise correlation matrix, and $\mathbf{R}_{\mathrm{cc}}(n,f)$ represents the noise-free SCM, which is given by:

$$\mathbf{R}_{\mathrm{cc}}(n,f) = \mathbf{E}(f)\mathbf{R}_{\mathrm{ss}}(n,f)\mathbf{E}^H(f). \tag{8}$$

Here,

$$\begin{aligned}
\mathbf{R}_{\mathrm{ss}}(n,f) &= E\left\{\mathbf{s}(n,f)\mathbf{s}^H(n,f)\right\}, \\
\mathbf{s}(n,f) &= [S_1(n,f), S_2(n,f), \ldots, S_K(n,f)]^T, \\
\mathbf{E}(f) &= [\mathbf{e}(f,\theta_1), \mathbf{e}(f,\theta_2), \ldots, \mathbf{e}(f,\theta_K)].
\end{aligned} \tag{9}$$

According to the spectral sparsity of speech signals, there exist TF bins where one source is dominant in energy and has significantly higher power than the others. The index of this source is:

$$\tilde{k}(n,f) = \arg \max_{k \in \{1,\ldots,K\}} E\left\{|S_k(n,f)|^2\right\}, \tag{10}$$

where $|\cdot|$ denotes the magnitude of complex number. In the following, the TF index of $\tilde{k}(n,f)$ is omitted for simplicity when it is utilized as a subscript. Ignoring the contribution of the sources with relatively lower intensity, the noise-free SCM for multiple sources is simplified to that for single source, which is given by:

$$\mathbf{R}_{\mathrm{cc}}(n,f) \approx E\left\{\left|S_{\tilde{k}}(n,f)\right|^2\right\}\mathbf{e}(f,\theta_{\tilde{k}})\mathbf{e}^H(f,\theta_{\tilde{k}}). \quad (11)$$

By multiplying $\mathbf{e}(f,\theta_{\tilde{k}})$ at both sides of Eq. (11), we have:

$$\mathbf{R}_{\mathrm{cc}}(n,f)\mathbf{e}(f,\theta_{\tilde{k}}) \approx E\left\{\left|S_{\tilde{k}}(n,f)\right|^2\right\}\left\|\mathbf{e}(f,\theta_{\tilde{k}})\right\|^2\mathbf{e}(f,\theta_{\tilde{k}}), \quad (12)$$

where $||\cdot||$ denotes the Euclidean norm. It is obvious that $\mathbf{R}_{\mathrm{cc}}(n,f)$ is approximately a rank-1 matrix, and $\mathbf{e}(f,\theta_{\tilde{k}})$ is its principal eigenvector, which spans the signal subspace of the $\tilde{k}(n,f)$th source.

B. Spatial Correlation Matrix

In order to estimate the SCM, a common way to approximate the expectation in Eq. (7) is averaging over different time frames. However, by examining the rank of this SCM, it is not able to effectively distinguish two cases, i.e., the SCM involving only the direct-path source and the SCM involving both the direct-path and reflections, due to the temporal correlation between direct-path and reflection signals. To overcome this, the expectation is approximated by further smoothing over frequency bands [32], [38]. Hence, the SCM for each TF bin is estimated by averaging over a predefined range of time frames and frequency bands. It is expressed as:

$$\hat{\mathbf{R}}_{\mathrm{xx}}(n,f) = \frac{1}{N_n N_f}\sum_{n'=n}^{n+N_n-1}\sum_{f'=f}^{f+N_f-1}\mathbf{x}(n',f')\mathbf{x}^H(n',f'), \quad (13)$$

where N_n and N_f denote the number of time frames and frequency bands used for approximating the SCM, respectively. N_f is set to 8 and N_n is set to 4. Since the average is computed over adjacent frequency bands, the resulting SCMs are highly redundant over frequency. Besides, the TF-wise processing based on $\hat{\mathbf{R}}_{\mathrm{xx}}(n,f)$ is computationally expensive due to the following eigen-analysis. To reach a tradeoff between computational complexity and frequency diversity, these TF bins are subsampled in frequency with a step of $N_f/2$, and only the subsampled TF bins are utilized in the following steps.

The SCM is contaminated with additive ambient noise. The power spectra of ambient noise are assumed to be identical for all microphones. For spatially white noise, the noise correlation matrix $\mathbf{R}_{\mathrm{vv}}(n,f)$ is a diagonal matrix. Such noise adds its power to all eigenvalues of $\mathbf{R}_{\mathrm{cc}}(n,f)$ uniformly without changing the principal eigenvector. Although the circumstance with spatially white noise is commonly considered for localization, diffuse noise field has been shown to be a more reasonable model for many practical noise fields [39]. Unlike spatially white noise, diffuse noise is highly spatially correlated especially for small arrays or at low frequency region [37]. The unknown spatial correlation significantly degrades the precision of the signal subspace estimation.

To remove the diffuse noise without destroying the signal subspace, the following two assumptions are helpful. One is that the spectral sparsity of speech signals makes $\mathbf{R}_{\mathrm{cc}}(n,f)$ low-rank. The other is that the diffuse noise has equivalent cross-spectra for microphone pairs spaced by the same distance, and consequently the geometrical symmetry of uniform circular arrays makes $\mathbf{R}_{\mathrm{vv}}(n,f)$ a circulant matrix [19], [20]. The circulant matrix can be diagonalized by a unitary discrete Fourier transform (DFT) matrix $\mathbf{P}$, i.e., $\mathbf{P}^H\mathbf{R}_{\mathrm{vv}}(n,f)\mathbf{P}$ is a diagonal matrix. According to Eq. (7), we obtain $\mathbf{P}^H\mathbf{R}_{\mathrm{vv}}(n,f)\mathbf{P} = \mathbf{P}^H\mathbf{R}_{\mathrm{xx}}(n,f)\mathbf{P} - \mathbf{P}^H\mathbf{R}_{\mathrm{cc}}(n,f)\mathbf{P}$. Hence, the noise correlation matrix only affects the diagonal entries of $\mathbf{P}^H\mathbf{R}_{\mathrm{xx}}(n,f)\mathbf{P}$, and the off-diagonal entries of $\mathbf{P}^H\mathbf{R}_{\mathrm{xx}}(n,f)\mathbf{P}$ is the same as that of $\mathbf{P}^H\mathbf{R}_{\mathrm{cc}}(n,f)\mathbf{P}$. Here, the unitary DFT matrix $\mathbf{P}$ is independent of $\mathbf{R}_{\mathrm{vv}}(n,f)$, and given by:

$$\mathbf{P} = \frac{1}{\sqrt{M}}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & \zeta^1 & \cdots & \zeta^{M-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \zeta^{M-1} & \cdots & \zeta^{(M-1)(M-1)} \end{bmatrix}, \quad (14)$$

with $\zeta = e^{-2\pi j/M}$.

Based on the aforementioned properties, $\mathbf{R}_{\mathrm{cc}}(n,f)$ can be estimated by [37]:

$$\hat{\mathbf{R}}_{\mathrm{cc}}(n,f) = \arg\min_{\mathbf{R}_{\mathrm{cc}}(n,f)} \frac{1}{2}\|\mathbf{P}\varphi\left\{\mathbf{P}^H\mathbf{R}_{\mathrm{cc}}(n,f)\mathbf{P}\right\}\mathbf{P}^H - \mathbf{P}\varphi\left\{\mathbf{P}^H\hat{\mathbf{R}}_{\mathrm{xx}}(n,f)\mathbf{P}\right\}\mathbf{P}^H\|_F^2 + \mu\left\|\mathbf{R}_{\mathrm{cc}}(n,f)\right\|_*, \quad (15)$$

where μ is a regularization factor, $\varphi\{\cdot\}$ denotes a function setting the diagonal entries to zeros, $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|_*$ denotes the trace norm. The Frobenius-norm term aims to recover $\mathbf{R}_{\mathrm{cc}}(n,f)$ from $\hat{\mathbf{R}}_{\mathrm{xx}}(n,f)$, i.e., to remove the influence of diffuse noise. The trace-norm regularization imposes a low-rank solution [40]. Under the constraint that $\mathbf{R}_{\mathrm{cc}}(n,f)$ is a positive semidefinite Hermitian matrix, this optimization problem can be solved by using the trace norm minimization algorithm presented in [37]. In addition to diffuse noise, other types of noise can also be removed as long as they have a circulant correlation matrix. Especially, the isotropic correlation matrix for spatially white noise is a special case of circulant matrix.

C. TF-Wise Spatial Spectrum

For each single source dominated TF bin, $\hat{\mathbf{R}}_{\mathrm{cc}}(n,f)$ can be approximated as a rank-1 matrix. Through the eigenvalue decomposition (EVD) of $\hat{\mathbf{R}}_{\mathrm{cc}}(n,f)$, the principal eigenvector $\mathbf{q}_1(n,f)$ that approximately spans signal subspace is obtained. The relation between $\mathbf{q}_1(n,f)$ and $\mathbf{e}(f,\theta_{\tilde{k}})$ can be expressed as [11]:

$$\mathbf{q}_1(n,f) \approx \frac{e^{-j\omega_f z}\mathbf{e}(f,\theta_{\tilde{k}})}{\left\|\mathbf{e}(f,\theta_{\tilde{k}})\right\|}, \quad (16)$$

where z is a real constant introduced by the complex EVD. The principal eigenvector approximates the steering vector with an extra complex constant gain, i.e., $\mathbf{q}_1(n,f)$ approximately points to $\theta_{\tilde{k}}$. Hence, for each single source dominated TF bin, the similarity between $\mathbf{q}_1(n,f)$ and $\mathbf{e}(f,\theta_{\tilde{k}})$ is taken to build the spatial

spectrum which can reflect the spatial possibility of source presence. Note that the signal subspace based spatial spectrum is estimated locally for each TF bin, which is referred to as TF-wise spatial spectrum.

In practice, spectral sparsity can be affected by multiple sources or reverberation, which decreases the similarity between $\mathbf{q}_1(n,f)$ and $\mathbf{e}(f,\theta_{\tilde{k}})$. To overcome this, a binary TF weight [32] is employed to select the single source dominated TF bins. The TF weight is computed as:

$$w_0(n,f)=\begin{cases}1, & \text{if } \dfrac{\lambda_1(n,f)}{\lambda_2(n,f)}>C_{\text{th}}\\ 0, & \text{otherwise}\end{cases}, \tag{17}$$

where $\lambda_1(n,f)$ and $\lambda_2(n,f)$ denote the largest and second-largest eigenvalues of $\hat{\mathbf{R}}_{\text{cc}}(n,f)$ respectively, and C_{th} denotes a predefined threshold. The threshold C_{th} is set as 3 referring to [26]. The TF bins with $w_0(n,f)=1$ are considered to be dominated by single source.

To build the TF-wise spatial spectrum, the 360-degree azimuth localization space is equally divided into D parts with each part corresponding to one candidate direction. The set of candidate directions is:

$$\mathbb{S}=\left\{0,1\times\frac{360}{D},\ldots,(D-1)\times\frac{360}{D}\right\}. \tag{18}$$

The number of candidate directions D is set to 360 in this work. For each single source dominated TF bin, the spatial spectrum value is obtained by calculating the similarity between the principal eigenvector and the steering vector related to each candidate direction. This is expressed as:

$$\rho(n,f,\theta)=\frac{\left|\mathbf{q}_1^H(n,f)\mathbf{e}(f,\theta)\right|}{\|\mathbf{e}(f,\theta)\|}, \tag{19}$$

where θ represents the candidate direction with $\theta\in\mathbb{S}$, and $\mathbf{e}(f,\theta)$ denotes the steering vector related to θ. By substituting Eq. (16) into Eq. (19), $\rho(n,f,\theta)$ can be approximated by:

$$\rho(n,f,\theta)\approx\frac{|\mathbf{e}(f,\theta_{\tilde{k}})^H\mathbf{e}(f,\theta)|}{\|\mathbf{e}(f,\theta_{\tilde{k}})\|\|\mathbf{e}(f,\theta)\|}. \tag{20}$$

The theoretical value of $\rho(n,f,\theta)$ in Eq. (20) is found to be irrelevant to the time frame index n. Besides, $\rho(n,f,\theta)$ ranges from 0 to 1, and reaches its maximum at the true direction of the local dominant source, i.e., $\theta=\theta_{\tilde{k}}$. Considering the 8-channel microphone array described in Section V, we plot the theoretical spatial spectrum as a function of the angular separation, namely $|\theta-\theta_{\tilde{k}}|$, for each frequency band in Fig. 2. Here, $\theta_{\tilde{k}}$ is set to 0°. It can be seen that the main peaks of spatial spectra are reached when $\theta=\theta_{\tilde{k}}$. The spatial spectrum slowly varies with an increasing angular separation for the low-frequency bands, and has more and more spurious peaks for the high-frequency bands. Since the spatial spectrum is usually used to reflect the possibility that one active sound source is present at each candidate direction, the spurious peaks far from the true direction should be low enough to avoid introducing ghost sources.

To increase the reliability of the spatial spectrum as an indicator of the source presence possibility, an exponential function

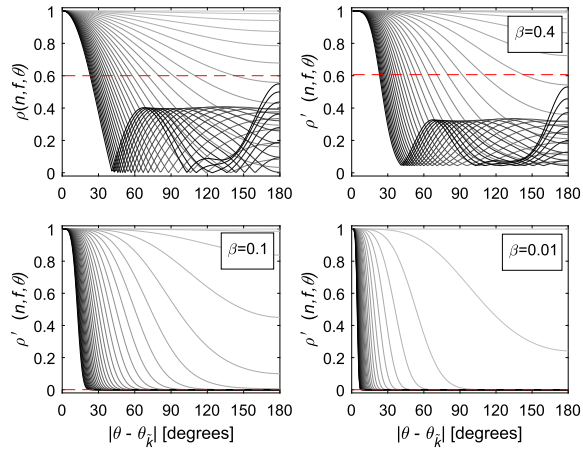


Fig. 2. Theoretical TF-wise spatial spectrum as a function of the angular separation $|\theta-\theta_{\tilde{k}}|$ ($\theta_{\tilde{k}}=0°$). Each curve corresponds to one frequency band, and darker gray represents higher frequency.

is employed to transform the original spatial spectrum into a domain where the higher spectrum value is maintained and the lower spectrum value is weakened. Specifically, the Gaussian function is adopted for the exponential transform. The improved TF-wise spatial spectrum is expressed as:

$$\rho'(n,f,\theta)=\exp\left\{-\frac{|1-\rho(n,f,\theta)|^2}{2\beta^2}\right\}, \tag{21}$$

where $\exp\{\cdot\}$ denotes the natural exponential function, and β is a positive adjustable variable which controls the attenuation degree of the lower spectrum value. Since the exponential transform monotonically maps $\rho(n,f,\theta)$ to $\rho'(n,f,\theta)$, the theoretical $\rho'(n,f,\theta)$ also ranges from 0 to 1, and reaches its maximum at the true direction of the local dominant source. The exponential transform indeed gives more significance to the high spatial spectrum values. The illustration of the theoretical spatial spectrum as a function of the angular separation $|\theta-\theta_{\tilde{k}}|$ is shown in Fig. 2. Three typical values of β are given, namely 0.4, 0.1 and 0.01. It can be seen that a smaller value of β results in a sharper main peak and a more significant attenuation of the spurious peaks. For a larger β (i.e., 0.4), there is no much difference between the spatial spectra with and without exponential transform. A very small β may bring some adverse effect in practice, i.e., the spatial spectrum value corresponding to the true source direction could be significantly suppressed when it is not very high. To obtain a tradeoff between the suppression of the spurious peaks and the preservation of true peaks, β is set to 0.1.

In the presence of reverberation or noise, the dominance of the spatial spectrum value corresponding to the true source direction becomes less prominent. For one TF bin, there are two cases: 1) The spatial spectrum value corresponding to the true source direction is still larger than the spurious peaks, for which the exponential transform is able to highlight the true peak. 2) The spurious peaks are comparable or larger than the spatial spectrum value corresponding to the true source direction, for which the exponential transform possibly harms the true peak. However, for this case, it is reasonable to expect that all the spatial spectrum values are not very high due to the low directivity

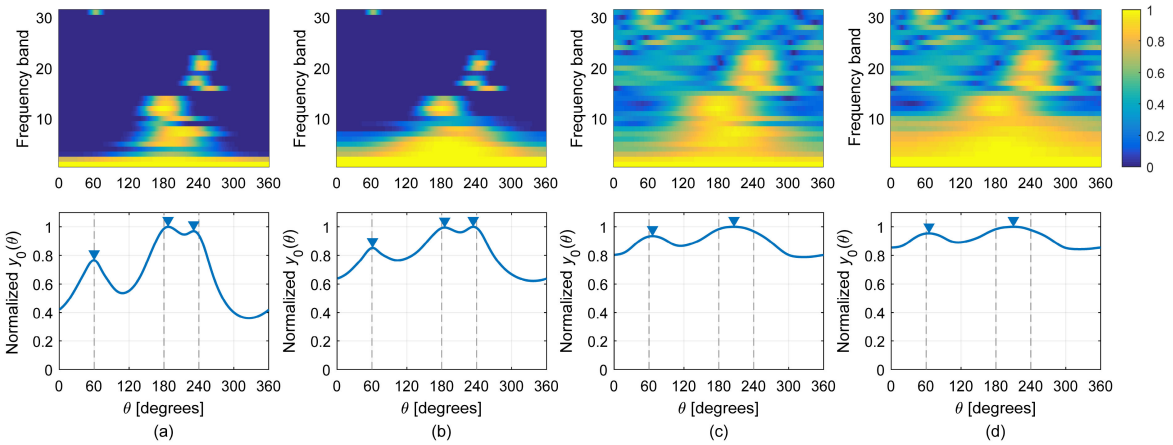


Fig. 3. Spatial spectrum for three sources at 60°, 180° and 240° in a simulated environment where $RT_{60} = 250$ ms and SNR = 5 dB (diffuse noise). Spectra are obtained (a) with SCM denoising and exponential transform, (b) with exponential transform but without SCM denoising, (c) with SCM denoising but without exponential transform, (d) without SCM denoising and exponential transform. First row: TF-wise spatial spectrum for each frequency band of one time frame. Second row: normalized global spatial spectrum obtained by using TF-wise spatial spectra in single source dominated TF bins.

of reverberation and noise. As a result, the entire spatial spectrum of this TF bin is suppressed relative to the one of other TF bins. In a way, exponential transform plays a role of selecting TF bins.

To illustrate the importance of SCM denoising and exponential transform to TF-wise spatial spectrum estimation, a three-source example is shown in Fig. 3. Note that the TF-wise spatial spectrum estimation without SCM denoising means the noisy SCM $\hat{\mathbf{R}}_{\mathrm{xx}}(n, f)$ is used, rather than the denoised $\hat{\mathbf{R}}_{\mathrm{cc}}(n, f)$. The first row depicts the TF-wise spatial spectra for different frequency bands. Comparing (a) with (b), it can be observed that the spatial spectra with SCM denosing show sharper peaks especially for low-frequency bands, and the same observation can be obtained from the comparison between (c) and (d). Comparing (a) with (c) (or (b) with (d)), we see that with exponential transform most spurious peaks and indistinctive peaks are removed from the spatial spectra, and the main peaks are preserved and become sharper. This phenomenon is more prominent at high-frequency bands. The second row depicts the normalized global spatial spectrum. The global spatial spectrum $y_0(\theta)$ is computed by summing $\rho'(n, f, \theta)$ over all single source dominated TF bins, and normalized by its maximum. Among the four figures, the normalized global spatial spectrum in (a) has the most accurate and sharpest peaks around the true directions.

IV. TF-Wise Spatial Spectrum Clustering

The global spatial spectrum is expected to present evident peaks only around true source directions as shown in Fig. 3(a). However, under adverse acoustic conditions, it is risky to count and localize sources based on the global spatial spectrum, since the peaks of the TF-wise spatial spectra would possibly be merged in the global spatial spectrum especially when sources are close to each other. This phenomenon is illustrated in Fig. 4. The interaction between sources from close directions causes that the number of peaks in the global spatial spectrum is smaller than the true number of sources, and the peak-associated directions deviate from the true source directions. Instead, the

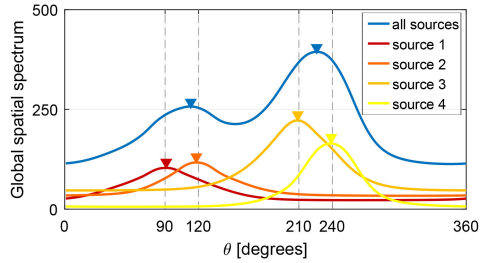


Fig. 4. Illustration of the impact of source interaction on the global spatial spectrum of all sources, in a simulated four-source environment where $RT_{60} = 250$ ms and SNR = 10 dB. The four sources are respectively located at 90°, 120°, 210° and 240°.

single-source global spatial spectra exhibit peaks around the true source locations with high precision. Note that, in Fig. 4, the global spatial spectrum for each source is obtained using the spatial spectrum-to-source assignment information, i.e., the TF-wise spatial spectra are assigned to the source direction (among 90°, 120°, 210° and 240°) closest to the local dominant source direction. This observation motivates us to classify the TF-wise spatial spectra into several clusters with each cluster corresponding to one source, and then estimate DOAs of sources from these clusters. The difficulty for spatial spectrum clustering lies in how to determine the number of clusters because the number of sources is usually unknown in realistic environment. Since the estimation of spatial spectrum clusters and their number are mutually affected, a method for jointly determining the two components is proposed.

The following two steps are often used in the proposed clustering method, 1) fusing a set of TF-wise spatial spectra, and 2) estimating DOA from the fused spectrum. Therefore, we first give the general formulations of them. The fused spectrum is computed by summing the TF-wise spatial spectra, which is formulated as:

$$y(\theta) = \sum_{n,f} w(n, f)\rho'(n, f, \theta), \tag{22}$$

where $w(n,f)$ denotes the TF weight whose binary value indicates the usage of the TF-wise spatial spectra (1 for selecting and 0 for discarding). When $w(n,f)=w_0(n,f)$, the fused spectrum is the global spatial spectrum for all sources, and represented by $y_0(\theta)$. When $w(n,f)$ associates with the TF bins dominated by one source, the fused spectrum is the single-source global spatial spectrum, and represented by $y_{k'}(\theta)$. The DOA associated with the maximum of the fused spectrum is:

$$\hat{\theta}=\arg\max_{\theta\in\mathbb{S}} y(\theta). \tag{23}$$

A. Iterative Source Detection Framework

In practice, sources are sparsely distributed at the candidate directions, and the number of sources is usually much smaller than the number of candidate directions. Based on the spatial sparsity of sources, an iterative source detection framework is designed. The details of each iteration are stated as follows.

New source detection: At the beginning of the kth iteration, $k-1$ sources have already been detected and the remaining global spatial spectrum $y_\Delta(\theta)$ has been computed. By applying Eq. (23) to $y_\Delta(\theta)$, the kth new potential source is detected and its DOA $\hat{\theta}_k$ is estimated. Note that for the first iteration, $y_\Delta(\theta)$ is initialized as $y_0(\theta)$.

Association adjustment: With the k detected sources, the association between sources and TF-wise spatial spectra (or TF bins) are adjusted. The association is indicated by $w_{k'}(n,f)$ with $k'\in\{1,2,\ldots,k\}$. Here, k' is the index of detected sources, and $w_{k'}(n,f)$ is the TF weight for the TF bins associated with the k'th source. According the adjusted association, the DOA estimates $\hat{\theta}_{k'}$ are also updated.

Remaining determination: Since the TF-wise spatial spectra associated with the detected sources are no longer used in the next iteration, the remaining TF-wise spatial spectra (or TF bins) need to be determined. The TF weight for the remaining TF bins is expressed as:

$$w_\Delta(n,f)=w_0(n,f)-\sum_{k'=1}^{k} w_{k'}(n,f). \tag{24}$$

The remaining global spatial spectrum is updated by fusing the TF-wise spatial spectra without the association of detected sources:

$$y_\Delta(\theta)=\sum_{n,f} w_\Delta(n,f)\rho'(n,f,\theta)=y_0(\theta)-\sum_{k'=1}^{k} y_{k'}(\theta). \tag{25}$$

By subtracting $y_{k'}(\theta)$ from the global spatial spectrum of all sources, the contribution of the detected sources is removed, and the influence of the k detected sources on the $(k+1)$th new source detection is reduced.

Stop criterion: In order to check whether the kth potential source is a real source or not, the change in the remaining TF bins and the remaining global spatial spectrum between the kth and $(k-1)$th iterations is regarded as the contribution of this source. When a real source is detected, $w_\Delta(n,f)$ or the peak value of $y_\Delta(\theta)$ shows significant changes between two adjacent external iterations. In contrast, when all the real sources have already been detected before the kth iteration, the changes caused by the new potential source is inapparent. Hence, the kth potential source is determined to be nonexistent if its contribution is indistinctive, i.e.,

$$\sum_{n,f}\left|w_\Delta(n,f)^{(k)}-w_\Delta(n,f)^{(k-1)}\right|<\eta_1\sum_{n,f} w_0(n,f), \tag{26}$$

or

$$\left|\max_{\theta\in\mathbb{S}} y_\Delta(\theta)^{(k)}-\max_{\theta\in\mathbb{S}} y_\Delta(\theta)^{(k-1)}\right|<\eta_2\sum_{n,f} w_0(n,f), \tag{27}$$

where the superscript $^{(k)}$ denotes the index of iteration, $\sum_{n,f} w_0(n,f)$ represents the number of single source dominated TF bins, and η_1 and η_2 are the predefined factors. In general, a large (or small) value of η_1 or η_2 leads to a small (or large) number of iterations. To achieve the best source counting performance, η_1 and η_2 are empirically set to 0.02 and 0.003, respectively. The maximum number of iterations is set to 10 to avoid a possible infinite number of iterations. If Eq. (26) or (27) is satisfied, which indicates that there are only $k-1$ sources, the iterative procedure will stop. Otherwise, the estimated number of the sources $\hat{K}$ is updated as k, the set of DOA estimates $\hat{\Theta}$ is updated as $\{\hat{\theta}_1,\ldots,\hat{\theta}_k\}$, and the iterative procedure continues to detect the next source.

The proposed iterative source detection framework can be seen as a generalization of the iterative contribution removal (ICR) algorithm presented in [34] and [35]. The generalization lies in that: 1) In the association adjustment process of each iteration, the ICR algorithm only adds the association between the newly detected source and TF bins, while the proposed source detection framework does not have this limitation and can adjust the association between all detected sources and TF bins. 2) The proposed framework can reestimate the DOAs of all temporarily detected sources according to the adjusted association after each new source detection, while the ICR algorithm cannot. The method in [35] can be regarded as an ICR algorithm followed by a DOA refining process, which is referred to as ICR+R in this work. Though the DOA reestimation is also provided by ICR+R, it is implemented by a DOA refining process after all sources are detected (i.e., after the ICR algorithm).

B. Proposed Clustering Method

The proposed clustering method is based on the iterative source detection framework. The three main parts, namely new source detection, association adjustment and remaining determination, are iteratively performed. In this framework, the association between sources and TF-wise spatial spectra is crucial to the performance. On the one hand, the DOA of each detected source is determined by the set of associated TF-wise spatial spectra. On the other hand, the TF-wise spatial spectra with no associated detected sources will be utilized for the next source detection. The dominance association between sources and TF-wise spatial spectra is considered for the association adjustment. With the estimated DOAs of all detected sources, namely $\{\hat{\theta}_1,\ldots,\hat{\theta}_k\}$, the association between the TF-wise spatial spectra and their

dominant sources is built in an iterative manner. To avoid confusion, the iteration for association adjustment is called internal iterative procedure while the iteration for source detection is called external iterative procedure. The details of each internal iteration are described as follows.

Spatial spectrum assignment: The TF-wise spatial spectra ought to be assigned to their respective dominant source. As shown in Fig. 2, $\rho'(n, f, \theta)$ for each TF bin approximately monotonically decreases with the increasing deviation of θ from $\theta_{\tilde{k}}$. The candidate direction with a larger spectrum value tends to be closer to the direction of local dominant source. Hence, to determine which source is possibly dominated in the (n, f)th TF bin, the largest spectrum value and corresponding source index are searched over all detected sources, which are formulated as:

$$S(n, f) = \max_{k' \in \{1,\ldots,k\}} \rho'(n, f, \hat{\theta}_{k'}), \tag{28}$$

$$\tilde{k}(n, f) = \arg \max_{k' \in \{1,\ldots,k\}} \rho'(n, f, \hat{\theta}_{k'}). \tag{29}$$

The largest spectrum value $S(n, f)$ is compared with a predefined similarity threshold S_{th} (ranging from 0 to 1) to further check whether the dominance of the $\tilde{k}(n, f)$th source is reliable. When $S(n, f) > S_{\text{th}}$, the spatial spectrum in the (n, f)th bin is thought to be dominated by the $\tilde{k}(n, f)$th source. Accordingly, the dominance association between the TF-wise spatial spectra and detected sources is represented by:

$$w_{k'}(n, f) = \begin{cases} 1, \text{ if } k' = \tilde{k}(n, f) \text{ and } S(n, f) > S_{\text{th}} \\ 0, \text{ otherwise} \end{cases}. \tag{30}$$

Regarding the value of $w_{k'}(n, f)$, 1 means that the spatial spectrum in the (n, f)th bin is assigned to the k'th source. The spatial spectra dominated by different sources are separated through the spatial spectrum assignment, and consequently the interaction between detected sources is reduced. Note that only spatial spectra in single source dominated TF bins are considered for the assignment. With a relatively larger S_{th}, a smaller number of spatial spectra are assigned to sources. The choice of S_{th} determines the spatial spectra that are employed to estimate the DOAs of detected sources, and the remaining spatial spectra that are utilized to detect the next potential source. Hence, S_{th} is crucial to the localization performance, and we discuss the setting of this parameter in Section V.

DOA estimation: Following Eq. (22), the single-source global spatial spectrum $y_{k'}(\theta)$ is computed for $k' \in \{1, \ldots, k\}$ using $w_{k'}(n, f)$. The DOA estimate of each source, namely $\hat{\theta}_{k'}$, is obtained by maximizing $y_{k'}(\theta)$ as Eq. (23). These DOA estimates are employed to guide the spatial spectrum assignment of the next internal iteration.

Stop criterion: The internal iteration terminates when the overall similarity almost remains the same between adjacent iterations. The overall similarity is obtained from all assigned spatial spectra, namely $\sum_{k'=1}^{k} y_{k'}(\hat{\theta}_{k'})$. The stop criterion is formulated as:

$$\left| \sum_{k'=1}^{k} y_{k'}(\hat{\theta}_{k'})^{(i)} - \sum_{k'=1}^{k} y_{k'}(\hat{\theta}_{k'})^{(i-1)} \right| < \delta, \tag{31}$$

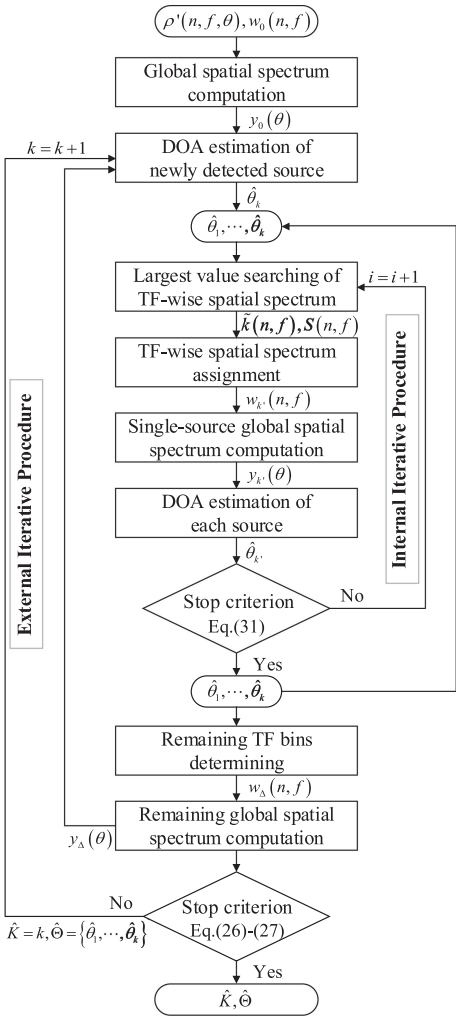


Fig. 5. Flowchart of the proposed TF-wise spatial spectrum clustering algorithm for joint source counting and localization. The external iterative procedure is aimed at counting sources one by one and the number of valid iterations corresponds to the number of source. The internal iterative procedure is designed to optimize the assignment of the spatial spectra to detected sources and refine the DOA estimates of sources.

where the superscript $^{(i)}$ denotes the index of internal iteration, and δ is a predefined threshold that controls the degree of association adjustment. To guarantee a sufficient adjustment of dominance association with a small number of iterations, δ is empirically set to 1. To avoid a possible infinite number of internal iterations, the maximum number of iterations is set to 5, since the iterative procedure usually terminates before 5 iterations in our experiments.

The proposed clustering method is summarized in Fig. 5, which includes external and internal iterative procedures. In each external iteration, one new cluster (source) is detected based on the remaining TF-wise spatial spectra, and then the TF-wise spatial spectra are reassigned to the already detected clusters through the internal iterative procedure.

V. Experiments and Discussions

Experiments are performed on simulated and real-world data using an 8-channel uniform circular array with radius $r = 5$ cm.

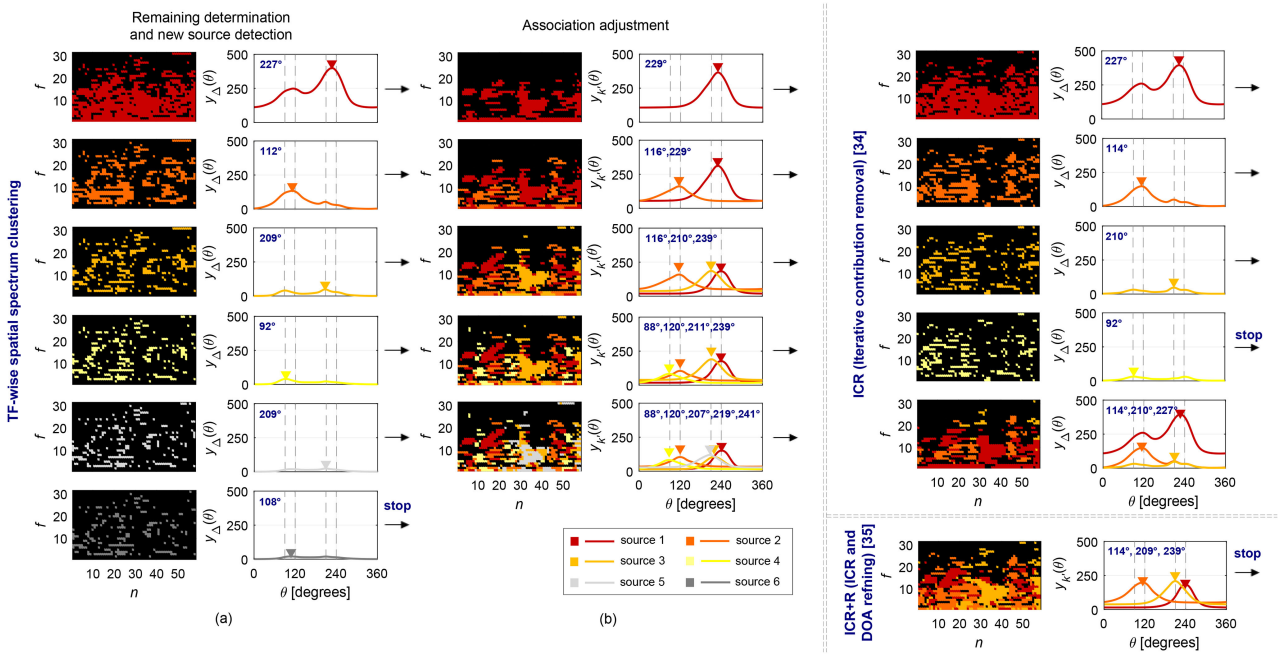


Fig. 6. Illustration of the SSC, ICR and ICR+R, in a simulated four-source environment where $RT_{60} = 250$ ms and SNR $= 10$ dB. The four sources are respectively located at 90°, 120°, 210° and 240°. The processes of our SSC method and ICR/ICR+R methods are illustrated on the left and right of the vertical dashed line, respectively. For the SSC method, (a) shows the remaining TF bins, and the new source detection procedure, and (b) shows the TF bin association adjustment procedure. The algorithm iterates between (a) and (b), namely from one row of (a) to the same row of (b) and to the next row of (a), and so on. More specifically, (a) **left** is the remaining (unassigned) TF bins, and (a) **right** is the remaining global spatial spectrum obtained by summing up the unassigned TF-wise spatial spectra, from which one new source is detected by searching the candidate direction with the largest spectrum value. (b) **left** shows the TF bins that are already assigned to one source, with one color for each source, and (b) **right** shows the global spatial spectrum for each source which is obtained by summing up the corresponding TF-wise spatial spectra. For the ICR method, the left column is the remaining TF bins, and the right column is the corresponding remaining global spatial spectrum. The last row summarizes the final output of ICR, with the left column corresponding to the removed TF bins associated with each source. The DOA refining process refines the final output of ICR. The left column is the TF bins assigned to each source, and the right column is the corresponding global spatial spectrum for each source.

In both cases, the array signals used for localization are with a duration of 1 s. The signal sampling rate is 16 kHz. The array signals are enframed by a window of 32 ms (512 samples) with a frame shift of 16 ms (256 samples). Only the frequency ranging from 0 to 4 kHz is considered for source localization.

The performance of multiple sound source localization is evaluated in two aspects, namely source counting and DOA estimation. The accuracy of source counting is measured with the recall rate, precision rate and F-score, which are respectively defined as:

$$R = \frac{\hat{K}_s}{K}, \quad P = \frac{\hat{K}_s}{\hat{K}}, \quad F = 2\frac{P \times R}{P + R} \tag{32}$$

where $\hat{K}_s$ is the number of successfully localized sources. The localization of each source is considered to be successful if the difference between the estimated DOA and the real DOA is smaller than a predefined threshold (namely 10° unless otherwise stated). The recall and precision rates reflect the missed and false detections of sources respectively, while the F-score reflects the overall counting performance. These three measures can also reflect the accuracy of DOA estimation to some extent. For all the three measures, a larger value indicates a better counting and localization performance. For the successfully localized sources, the accuracy of DOA estimation is measured with the mean absolute error (MAE) between the estimated DOAs and the ground-truth DOAs.

To evaluate the effectiveness of the proposed TF-wise spatial spectrum clustering (SSC) method, three sound source counting and localization methods are taken for comparison, namely direct peak counting (DC) [11], ICR [34], and ICR+R [35]. For fair comparison, all the methods use the TF-wise spatial spectrum proposed in this work. They are based on the combination of spatial spectrum and TF processing. The DC method estimates the number and DOAs of sources by searching the significant peaks of $y_0(\theta)$ whose values are larger than a predefined threshold. The ICR, ICR+R and our method count and localize each source successively based on the association between the TF bins and sources. The ICR method in [34] is adapted to the use of the TF-wise spatial spectra, and it is identical to the counting part of ICR+R in practical implementation.

A. An Example of Spatial Spectrum Clustering

To illustrate how the SSC, ICR and ICR+R methods work on the TF-wise spatial spectra, an example is shown in Fig. 6 using the same data as in Fig. 4. At the first of the SSC method, the spatial spectra in single source dominated TF bins are used to estimate the DOA of the first source (227°). Then, the association between spectra and the first source is adjusted, and only the spectra assigned to the first source are used to reestimate its DOA (229°). The remaining TF bins, which excludes the already assigned TF bins, are utilized to localize the second source

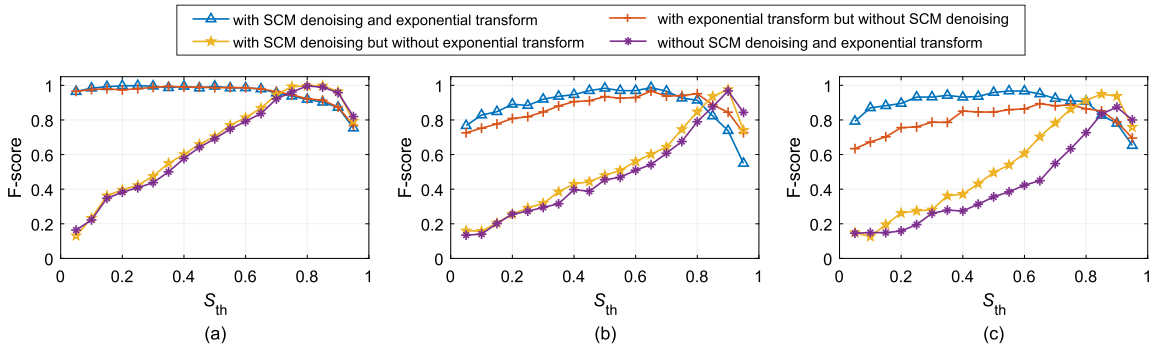


Fig. 7. Performance (F-score) versus similarity threshold S_{th} for different settings of TF-wise spatial spectrum estimation method. Results are obtained in three simulated four-source scenarios with (a) $\text{RT}_{60} = 400$ ms and SNR $= 25$ dB (diffuse noise), (b) $\text{RT}_{60} = 200$ ms and SNR $= 0$ dB (spatially white Gaussian noise), (c) $\text{RT}_{60} = 200$ ms and SNR $= 0$ dB (diffuse noise).

(112°). Considering the currently available DOA estimates (229° and 112°), the association between the TF-wise spatial spectra and the two detected sources is adjusted, and the DOA of the two sources are reesimated (229° and 116°). Similar actions are performed to localize the following sources. When sources are gradually detected, the change in $w_{\Delta}(n, f)$ or $y_{\Delta}(\theta)$ between iterations becomes smaller. The iterative procedure stops when the change introduced by the fifth source is found to be insignificant. The final DOA estimates of the four sources are obtained in the association adjustment before the fifth source detection (88°, 120°, 211° and 239°). It can be seen that the data processing of the proposed method is indeed a clustering of the TF-wise spatial spectra according to their dominant sources. For the ICR method, the iteration stops when the change introduced by the fourth source is indistinctive. The DOA estimates provided by ICR are 114°, 210° and 227°. With the DOA refining, the DOA estimates are optimized and become 114°, 209° and 239°. Since the TF-wise spatial spectra are iteratively removed and the remaining spectra are used for new source detection in the ICR method, there still remains the peak merging problem caused by the interaction between sources, especially for the detection of the first few sources. The peak merging problem results in a cumulative error in estimating the DOAs and determining the remaining TF-wise spatial spectra used for next source detection, which further affects the overall counting and localization results. In the ICR+R method, though the DOA refining process helps to optimize the DOA estimates obtained from the ICR algorithm, it fails to change the estimated number of sources. In contrast to the ICR and ICR+R methods, the proposed method iteratively adjusts the association between TF-wise spatial spectra and sources after each source detection, and aims at clustering the TF-wise spatial spectra. In this way, the peak merging problem is suppressed and better localization performance is achieved.

B. Evaluation With Simulated Data

Simulated data is obtained from a room with the size of 6 m × 5 m × 3 m. The microphone array is placed at the center of the room. Sound sources are located in same horizontal plane as the array with a array-to-source distance of 1.5 m. The directions of sources are randomly set in the range from 0° to 360° with an interval of 5°, and the minimum angular separation between sources is set to 30° unless otherwise stated. The speech recordings from the TIMIT database [41] are taken as the source signals. To generate the room impulse responds (RIRs), the image method [42] implemented using toolbox [43] is adopted. The microphone signals are obtained by convolving the clean source signals with the generated RIRs. To control the signal-to-noise ratio (SNR), additive ambient noise is properly scaled and added to each microphone signal. Here, diffuse noise [39] is utilized unless otherwise stated.

1) Influence of S_{th}: We investigate the effect of the parameter S_{th} on the localization performance. In Fig. 7, the F-score obtained by the TF-wise spatial spectrum clustering algorithm is depicted as a function of S_{th} for different settings of SCM denosing and exponential transform. The simulations are performed in the four-source scenarios with different acoustic conditions: $\text{RT}_{60} = 400$ ms and SNR $= 25$ dB (diffuse noise) in (a), $\text{RT}_{60} = 200$ ms and SNR $= 0$ dB (spatially white Gaussian noise) in (b), and $\text{RT}_{60} = 200$ ms and SNR $= 0$ dB (diffuse noise) in (c). The presented results are an average of 100 instances with different source directions.

When exponential transform is applied, the proposed SSC method performs consistently with increasing S_{th} under different acoustic conditions. With increasing S_{th}, the F-score varies slowly when $S_{\text{th}} < 0.8$, and drops sharply when $S_{\text{th}} > 0.8$. Considering different acoustic conditions, the global optimal performance can be achieved when S_{th} ranges from 0.4 to 0.7. When exponential transform is not applied, the F-score increases until S_{th} reaches about 0.8, and decreases when $S_{\text{th}} > 0.85$. For this case, the global optimal performance can be obtained when S_{th} varies from 0.85 to 0.9. The reason for that an improper setting of S_{th} could lead to a bad localization performance is stated as follows. For extremely small S_{th} (approximate to 0), too many TF bins are assigned to the early detected sources, and the contribution of a new source is possibly inapparent. For extremely large S_{th} (approximate to 1), only a small number of TF bins are assigned to each source. Both cases can make the external iterative procedure stop before all real sources are detected. Besides, when exponential transform is not applied, there are several spurious peaks in TF-wise spatial spectra. Due to the ambiguity caused by the spurious peaks, using relatively smaller S_{th} will introduce a large error to the assignment of the

TABLE I
PERFORMANCE (F-SCORE) FOR DIFFERENT SETTINGS OF THE TF-WISE SPATIAL SPECTRUM ESTIMATION METHOD, AND FOR DIFFERENT LOCALIZATION METHODS

RT_{60}	SNR	SSC				ICR+R				ICR				DC			
		E1	E2	E3	E4	E1	E2	E3	E4	E1	E2	E3	E4	E1	E2	E3	E4
200ms	25dB	**1.00**	**1.00**	**1.00**	0.99	**0.98**	**0.98**	0.88	0.88	**0.97**	0.96	0.78	0.76	**0.87**	0.86	0.45	0.46
200ms	5dB	**0.99**	0.97	0.97	0.96	0.94	**0.95**	0.83	0.79	**0.86**	0.83	0.66	0.64	**0.61**	0.56	0.23	0.24
200ms	0dB	**0.96**	0.89	0.94	0.87	**0.91**	0.78	0.80	0.64	**0.78**	0.66	0.51	0.48	**0.47**	0.35	0.19	0.17
400ms	25dB	**0.99**	0.98	**0.99**	0.96	**0.97**	0.96	0.87	0.83	**0.95**	0.94	0.72	0.72	**0.80**	0.80	0.37	0.37

E1: with SCM denoising and exponential transform
E2: with exponential transform but without SCM denoising
E3: with SCM denoising but without exponential transform
E4: without SCM denoising and exponential transform

TF-wise spatial spectra to detected sources, which can further affects the stop of external iterative procedure.

The parameter $S_{\rm th}$ largely affects the localization performance. The values of $S_{\rm th}$ that achieve the optimal localization performance are actually quite consistent under different acoustic conditions. It can be observed from Fig. 7 that the optimal performance for different acoustic conditions is achieved with the similar $S_{\rm th}$ setting. Accordingly, $S_{\rm th}$ is set to 0.5 for SSC when using the TF-wise spatial spectrum estimation method with SCM denoising and exponential transform, which is found to be the optimal value under various acoustic conditions in terms of noise levels, reverberation times and number of sources. Similarly, based on some preliminary experiments, the values of $S_{\rm th}$ are set to 0.55 for both ICR+R and ICR when using the TF-wise spatial spectrum estimation method with SCM denoising and exponential transform.

The effectiveness of SCM denoising and exponential transform to localization performance is also verified in Fig. 7. When the diffuse noise is significant, i.e., in Fig. 7(c), the method with SCM denosing achieves better optimal performance compared with that without SCM denosing. This verifies that the SCM denoising can improve the localization robustness against diffuse noise. By applying exponential transform, the range of $S_{\rm th}$ that achieves the optimal performance is enlarged, and the setting of $S_{\rm th}$ is facilitated accordingly. This is mainly attributed to the enlarged gap between the large and small spectrum values by the exponential transform. To further confirm the effectiveness of SCM denoising and exponential transform, a more detailed comparison is shown in Table I. Note that the presented F-scores represent the optimal performance of each setting after parameter selection. For the DC method, the threshold for detecting significant peaks is set to 0.2 times the maximum value of $y_0(\theta)$, which achieves a good tradeoff for DC under different acoustic conditions. It can be seen that, with the increasing of the noise level, the performance improvement caused by SCM denoising becomes more prominent. The performance is also improved by the exponential transform, which is more obvious for ICR+R, ICR and DC. For each localization method, the TF-wise spatial spectrum estimation with both SCM denoising and exponential transform outperforms the other three settings. Besides, SSC tends to outperform the other localization methods with the same spectrum setting, even when SCM denoising and exponential transform are not adopted. It confirms that the advantage of SSC over the other methods is not due to the use of SCM denoising and exponential transform.

2) Performance Comparison: The comparison of the proposed method with ICR+R, ICR and DC is carried out in the environments with different SNRs, reverberation times and numbers of active sources. The number of active sound sources is in the range 2 to 5. Fig. 8 shows the resulting F-scores. It can be observed that the SSC method outperforms the other three methods for almost all conditions. The superiority of the SSC is more prominent when the acoustic condition worsens. For example, under the five-source case with $RT_{60} = 250$ ms and SNR $=$ 0 dB, the F-score of SSC is about 0.84, and exceeds ICR+R, ICR and DC by about 0.15, 0.25 and 0.5 respectively. As the number of sources, the SNR or the RT_{60} increases, the performance of all the methods degrades due to increased distortion caused by acoustic interference. The performance degradation of SSC is relatively smaller among the four methods, indicating that it is more robust against acoustic interference.

Fig. 9 depicts a comprehensive performance comparison in terms of recall rate, precision rate, F-score and MAE of all the four methods. The performance is obtained for different numbers of sources in the simulated environment where $RT_{60} = 250$ ms and SNR $= 10$ dB. From the recall rate, precision rate and F-score, it can be inferred that the SSC method achieves the least missed and false source detections. ICR+R and ICR perform worse than SSC in both aspects, especially when the number of sources increases. The DC method provides the worst performance for source detection, and tends to underestimate the number of sources. For the accuracy of DOA estimation, SSC achieves a comparable MAE to ICR+R, which is lower than ICR and DC. The performance of both source counting and DOA estimation degrades with a increasing number of sources.

An example is shown in Fig. 10 to illustrate the behavior of the tested methods. In this instance, RT_{60} is 250 ms and SNR is 10 dB. The static speech sources are located at 60°, 150°, 180°, 230° and 310° around the array. As indicated by the gray dots in Fig. 10, during the twenty-second test period, the number of active sources increases from two to five and then decreases to two again. The number and DOAs of sources are jointly estimated for every one-second sensor recording with a step of 0.1 s. Since one-second history recording is used for each estimation, the DOA estimation is delayed for certain period after sources become active or inactive. For sources located at 60° and 310° which are far away from other sources, the performance of all the four methods is comparable and acceptable. For closely located sources at 150°, 180° and 230°, the proposed SSC method shows a significant advantage over the compared methods.

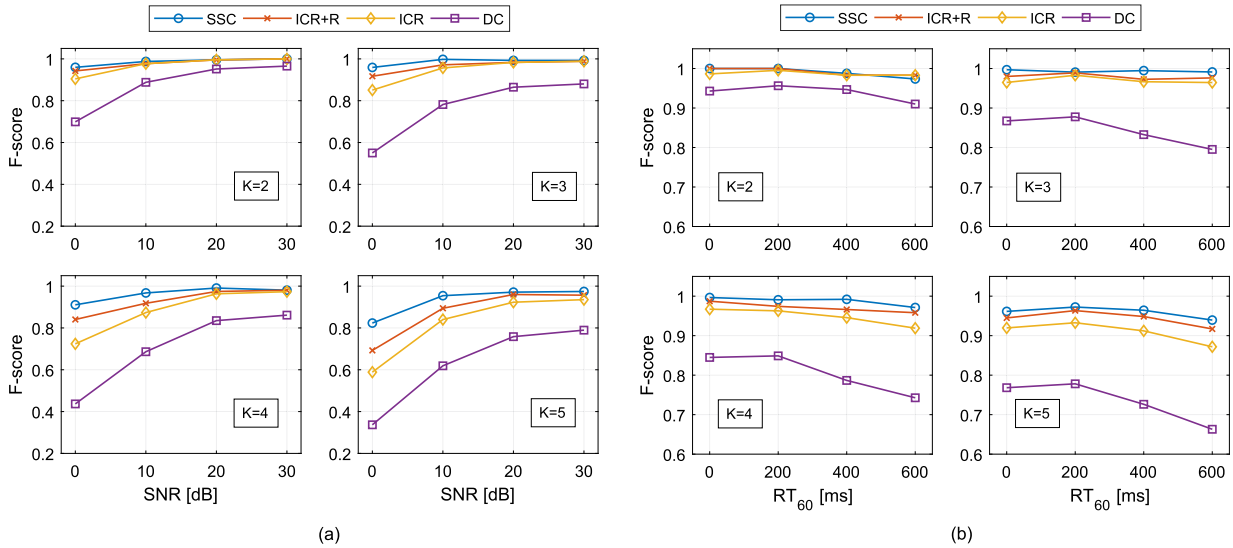


Fig. 8. Performance (F-score) comparison of different methods in the simulated scenarios with (a) different SNRs and numbers of sources ($RT_{60} = 250$ ms), (b) different RT_{60}s and numbers of sources (SNR $= 20$ dB).

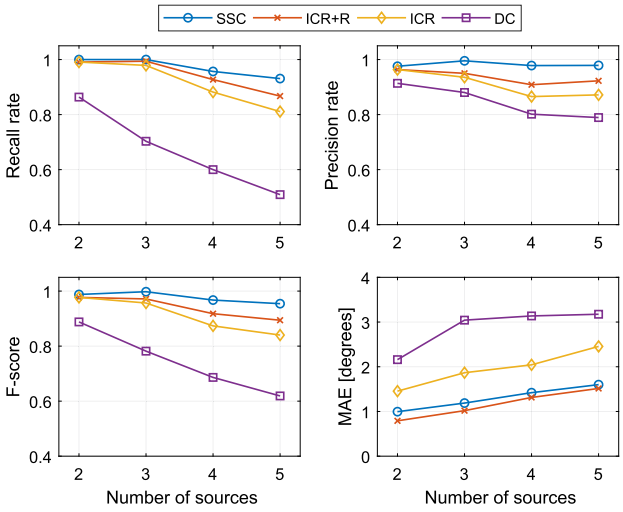


Fig. 9. Detailed performance (recall rate, precision rate, F-score and MAE) comparison of different methods in the simulated scenarios with different number of sources ($RT_{60} = 250$ ms, SNR $= 10$ dB).

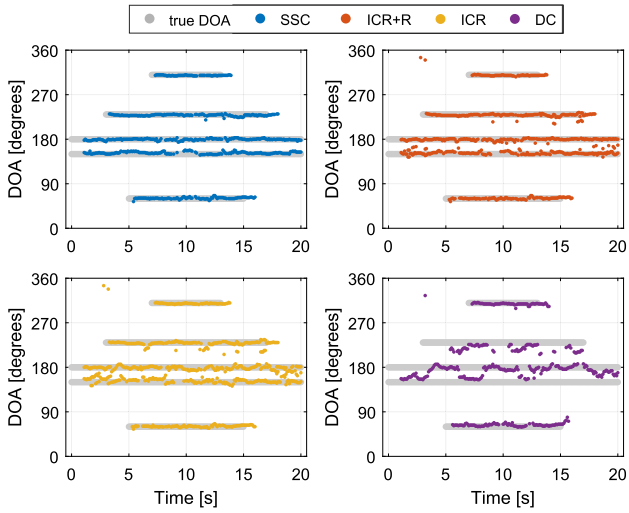


Fig. 10. Illustration of joint source counting and localization using SSC, ICR+R, ICR and DC respectively, in a simulated environment where $RT_{60} = 250$ ms and SNR $= 10$ dB.

Generally, the SSC method achieves the least missed and false source detections and the most accurate DOA estimation, which manifests the superiority of the proposed method for joint source counting and localization. The ICR+R method provides slightly better results in terms of the accuracy of DOA estimation than ICR, which is attributed to the DOA refining process. The DC method performs the worst, and the significant performance gap between scattered sources and gathered sources confirms that DC is sensitive to the angular separation between sources.

3) Spatial Resolution: To investigate the spatial resolution of each method, namely the minimum angular separation between two sources when they can be successfully counted and localized, the F-score is plotted as a function of the angular separation from 10° to 180° with a step of 10° in Fig. 11. The localization of each source is considered to be successful if the estimated DOA deviates no more than 5° from the real DOA. The number of sources is two, RT_{60} is 250 ms and SNR is 15 dB. To produce various instances, the two sources are rotated with 7 different angles (from 0° to 90° with a step of 15°) from their initial locations, and for each rotation 10 different source signals are set. With regard to each angular separation, the presented results are averaged over these 70 instances.

It can be observed that SSC fails to successfully count and localize sources with an angular separation of 10°, since the contribution of two closely located sources cannot be effectively separated. As the angular separation increases from 10° to 20°, the F-score for SSC increases sharply. Accordingly, we can say that the spatial resolution of SSC with the given microphone array in this environment is about 20°. The proposed SSC method achieves better spatial resolution than ICR+R, ICR

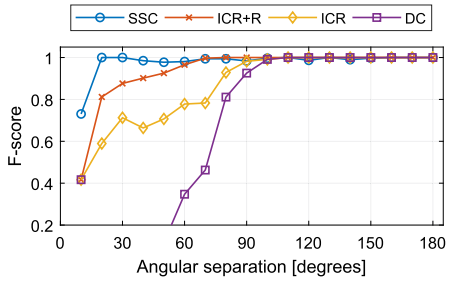


Fig. 11. Performance (F-score) versus angular separations for different methods. Results are obtained in a simulated two-source scenario where $RT_{60} = 250$ ms and SNR $= 15$ dB.

TABLE II
COMPUTATION TIME [S] OF DIFFERENT METHODS

Method	K=2	K=3	K=4	K=5
SSC	1.23	1.33	1.43	1.55
ICR+R	1.13	1.14	1.15	1.17
ICR	1.08	1.08	1.08	1.08
DC	1.06	1.05	1.04	1.04

and DC especially when the angular separation between sources is smaller than 60°. The advantage is mainly attributed to the reduced source interaction by the proposed iterative source detection framework and the dominance association adjustment.

4) Computational Complexity: The proposed method consists of three parts, namely SCM computation (see Section III-B), TF-wise spatial spectrum estimation (see Section III-C) and TF-wise spatial spectrum clustering (see Section IV). The computational complexity of SCM computation is dominated by the diffuse noise removal that consists of an iterative optimization procedure for each TF bin. Hence, the main complexity of this part depends on the number of TF bins and the number of optimization iterations. The computational complexity of the TF-wise spatial spectrum estimation is related to the number of TF bins and the number of candidate directions. Since the clustering part is implemented based on the selected TF-wise spatial spectra, its computational complexity depends on the number of single source dominated TF bins, the number of candidate directions, and the number of external and internal iterations. In this work, we implement the ICR+R, ICR and DC methods using the same TF-wise spatial spectra as in the proposed method, hence their computational complexity only differs in the localization procedure.

We implement SSC, ICR+R, ICR and DC in Matlab, and run them on an Intel CPU. The simulated microphone signals are generated with $RT_{60} = 250$ ms and SNR $= 20$ dB. The number of sources is set from 2 to 5. For each source-number configuration, 100 localization instances are performed, with a signal duration of 1 s for each instance. Table II reports the computation time averaged over all localization instances for each source-number configuration. The proposed method has a slightly higher computation time than the other methods. With the increasing number of sources, the computation time of the proposed method grows a little, while the computation time of the other three methods keeps almost constant. Table III lists

TABLE III
COMPUTATION TIME [S] OF DIFFERENT PARTS OF THE PROPOSED METHOD

Part	K=2	K=3	K=4	K=5
SCM computation	0.77	0.76	0.75	0.75
TF-wise spatial spectrum estimation	0.29	0.29	0.29	0.28
TF-wise spatial spectrum clustering	0.17	0.28	0.39	0.52
overall	1.23	1.33	1.43	1.55

TABLE IV
PERFORMANCE (F-SCORE) COMPARISON OF DIFFERENT METHODS IN THE REALISTIC ENVIRONMENT

Method	K=2	K=3	K=4	K=5
SSC	0.98	0.97	0.95	0.88
ICR+R	0.97	0.96	0.93	0.86
ICR	0.97	0.96	0.91	0.82
DC	0.95	0.88	0.80	0.69

the computation time of each part of the proposed method. The SCM computation represents a large proportion in the overall computation time. Along with the increasing of the number of sources, the computation time of SCM computation and TF-wise spatial spectrum estimation remains almost constant, while the computation time of clustering part grows.

C. Evaluation With Real-World Data

Real-world data is collected in a normal office room which has approximately identical dimensions and microphone placement to that for simulated data. The RT_{60} of this room is about 400 ms and the SNR is about 10 dB. We use eight Shure SM93 omnidirectional microphones[1] to constitute the uniform circular array. The microphone array together with a TASCAM US-16 × 8 USB sound card[2] is used for audio recording. To create active sound sources, the audio files selected from the TIMIT database are played by a loudspeaker. The loudspeaker is located around the array with a distance of 1.5 m, in the same horizontal plane as the array. Sources from the directions 48°, 94°, 151°, 180°, 241°, 302°, 359° are recorded separately, and K out of the 7 single-source recordings are selected and superimposed to form multi-source sensor signals. Hence, for $K = 2$, 3, 4 and 5, we respectively have 21, 35, 35 and 21 source direction combinations. For each source direction combination, 10 instances with different source signals are used for performance evaluation.

Table IV shows the performance comparison in terms of F-score of the four localization methods. It can be observed that the proposed SSC method achieves the highest F-score under all conditions, which demonstrates the effectiveness of the proposed iterative source detection framework and the dominance association adjustment in realistic scenarios. ICR+R and ICR obtain relatively worse performance. The DC method performs the worst and its performance degrades significantly with increasing number of sources. Overall, the performance measures

[1]http://www.shure.com/americas/products/microphones/sm/sm93-lavalier-microphone

[2]http://tascam.cn/product/us-16x08/

in the realistic environment are almost consistent with the results obtained on the simulated data.

VI. Conclusion

This paper proposes a TF-wise spatial spectrum clustering method to robustly count and localize multiple sound sources in adverse acoustic environments. The SCM denoising algorithm is adopted to improve the robustness of spatial spectrum against diffuse noise. The exponential transform is designed to enlarge the gap between large and small spectrum values, which can increase the reliability of spatial spectrum as an indicator of the source presence possibility. The two improvements are verified to guarantee a favorable performance under different acoustic conditions. The TF-wise spatial spectrum clustering algorithm is proposed for joint source counting and localization. The characteristics of the proposed algorithm lie in: 1) The spatial spectra are gradually clustered without using a priori about the number of sources, which is realized by the iterative source detection framework. 2) The spatial spectra in each cluster are dominated by the same source, which is ensured by the dominance association adjustment. Both of these two aspects help to reduce the interaction between sources, and hence superior performance can be achieved even when sources are close to each other. Experiments conducted on both simulated and real-world data demonstrate that the proposed method can achieve significant improvement in joint source counting and localization when compared with several other methods. In addition, the proposed method shows superior adaptability under various of conditions including different levels of noise and reverberation, numbers of sources and angular separations between sources.

In this work, the association adjustment is crucial to the localization performance. However, a small angular separation of sources or a large number of sources may increase the ambiguity of the association between spatial spectra and sources. In this case, other acoustic features (e.g., pitch) can be considered to correct the ambiguous association. Since this work only deals with the static sources, we may try to reduce the complexity of the proposed method and extend it for real-time localization of moving sources in the future.

References

[1] H. Do and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 125–128.

[2] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[3] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[4] T. N. T. Nguyen, S. Zhao, and D. L. Jones, "Robust DOA estimation of multiple speech sources," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 2287–2291.

[5] 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.

[6] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, Oct. 2013.

[7] J. Zhang and H. Liu, "Robust acoustic localization via time-delay compensation and interaural matching filter," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4771–4783, Sep. 2015.

[8] X. Li and H. Liu, "Sound source localization for HRI using FOC-based time difference feature and spatial grid matching," *IEEE Trans. Cybern.*, vol. 43, no. 4, pp. 1199–1212, Aug. 2013.

[9] 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.

[10] X. Xiao, S. Zhao, T. N. T. Nguyen, D. L. Jones, E. S. Chng, and H. Li, "An expectation-maximization eigenvector clustering approach to direction of arrival estimation of multiple speech sources," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 6330–6334.

[11] D. Ying, R. Zhou, J. Li, and Y. Yan, "Window-dominant signal subspace methods for multiple short-term speech source localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 731–744, Apr. 2017.

[12] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer, 2008.

[13] S.-C. Lee, B.-W. Chen, J.-F. Wang, M.-J. Liao, and W. Ji, "Subspace-based DOA with linear phase approximation and frequency bin selection preprocessing for interactive robots in noisy environments," *Comput. Speech Lang.*, vol. 34, no. 1, pp. 113–128, 2015.

[14] C. Xu, X. Xiao, S. Sun, W. Rao, E. S. Chng, and H. Li, "Weighted spatial covariance matrix estimation for music based TDOA estimation of speech source," in *Proc. INTERSPEECH*, 2017, pp. 1894–1898.

[15] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1327–1339, May 2007.

[16] M. Souden, J. Benesty, and S. Affeso, "Broadband source localization from an eigenanalysis perspective," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1575–1587, Aug. 2010.

[17] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–19, 2006.

[18] W. Xue and W. Liu, "Direction of arrival estimation based on subband weighting for noisy conditions," in *Proc. INTERSPEECH*, 2012, pp. 142–145.

[19] H. Shimizu, N. Ono, K. Matsumoto, and S. Sagayama, "Isotropic noise suppression in the power spectrum domain by symmetric microphone arrays," in *Proc. IEEE Work. Appl. Signal Process. Audio Acoust.*, 2007, pp. 54–57.

[20] N. Ito, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Crystal-MUSIC: Accurate localization of multiple sources in diffuse noise environments using crystal-shaped microphone arrays," in *Proc. Int. Conf. Latent Variable Anal., Signal Separation*, 2010, pp. 81–88.

[21] R. Y. Litovskya, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 1633–1654, 1999.

[22] J. Huang, N. Ohnishi, and N. Sugie, "Sound localization in reverberant environment based on the model of the precedence effect," *IEEE Trans. Instrum. Meas.*, vol. 46, no. 4, pp. 842–846, Aug. 1997.

[23] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 320–324.

[24] C. Pang, H. Liu, J. Zhang, and X. Li, "Binaural sound localization based on reverberation weighting and generalized parametric mapping," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 8, pp. 1618–1632, Aug. 2017.

[25] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 2136–2147, 2008.

[26] B. Rafaely and D. Kolossa, "Speaker localization in reverberant rooms based on direct path dominance test statistics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 6120–6124.

[27] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Process.*, vol. 92, no. 8, pp. 1950–1960, 2012.

[28] Z. Huang, G. Zhan, D. Ying, and Y. Yan, "Robust multiple speech source localization using time delay histogram," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 3191–3195.

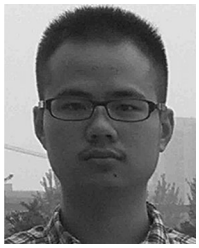
[29] S. HaJezi, A. H. Moore, and P. A. Naylor, "Multiple source localization using estimation consistency in the time-frequency domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 516–520.

[30] A. M. Torres, M. Cobos, B. Pueo, and J. J. Lopez, "Robust acoustic source localization based on modal beamforming and time-frequency processing using circular microphone arrays," *J. Acoust. Soc. Amer.*, vol. 132, no. 3, pp. 1511–1520, 2012.
[31] Z. E. Chami, A. Guerin, A. Pham, and C. Serviere, "A phase-based dual microphone method to count and locate audio sources in reverberant rooms," in *Proc. IEEE Work. Appl. Signal Process. Audio Acoust.*, 2009, pp. 209–212.
[32] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.
[33] X. Li, L. Girin, R. Horaud, and S. Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1997–2012, Oct. 2017.
[34] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1079–1093, Jun. 2016.
[35] B. Yang, H. Liu, and C. Pang, "Multiple sound source counting and localization based on spatial principal eigenvector," in *Proc. INTERSPEECH*, 2017, pp. 1924–1928.
[36] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[37] N. Ito, E. Vincent, N. Ono, and S. Sagayama, "Robust estimation of directions-of-arrival in diffuse noise based on matrix-space sparsity," INRIA, Le Chesnay, France, Res. Rep. RR-8120, 2012.
[38] H. Wang, C. C. Li, and J. X. Zhu, "High-resolution direction finding in the presence of multipath: A frequency-domain smoothing approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1987, pp. 2276–2279.
[39] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3464–3470, 2007.
[40] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific J. Optim.*, vol. 6, pp. 615–640, 2010.
[41] J. S. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, PA, USA, 1993. [Online]. Available: https://catalog.ldc.upenn.edu/ldc93s1
[42] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[43] E. A. P. Habets, "RIR generator." [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-gen erator

Bing Yang received the B.E. degree in automation from the University of Science and Technology Beijing, Beijing, China, in 2015. She is currently working toward the Ph.D. degree with the School of Electronics Engineering and Computer Science, Peking University, Beijing, China. Her current research interests include multimicrophone speech and audio processing for sound source localization and tracking.

Hong Liu received the Ph.D. degree in mechanical electronics and automation from Harbin Institute of Technology, Harbin, China, in 1996.

He is currently a Full Professor with the School of Electronics Engineering and Computer Science, Peking University (PKU), Beijing, China. He is also the Director of Open Lab on Human Robot Interaction, PKU. He has authored or coauthored more than 150 papers. His research interests include computer vision and robotics, image processing, and pattern recognition. He was the recipient of the Chinese National Aerospace Award, the Wu Wenjun Award on Artificial Intelligence, the Excellence Teaching Award, and the Candidates of Top Ten Outstanding Professors in PKU. He has been selected as the Chinese Innovation Leading Talent supported by the National High-level Talents Special Support Plan since 2013. He is the Vice President of the Chinese Association for Artificial Intelligent (CAAI), and the Vice Chair of the Intelligent Robotics Society, CAAI. He was keynote speakers, Co-Chairs, Session Chairs, or PC members of many important international conferences, such as IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE International Conference on Robotics and Biomimetics, IEEE International Conference on Systems, Man, and Cybernetics, and International Conference on Intelligent Information Hiding and Multimedia Signal Processing. He is also reviewers for many international journals such as *Pattern Recognition*, IEEE TRANSACTIONS ON SIGNAL PROCESSING, and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.



Cheng Pang received the B.E. degree in mechatronic engineering, in 2013. He is currently working toward the Ph.D. degree with the School of Electronics Engineering and Computer Science, Peking University, Beijing, China. His current research interests include speech and audio signal processing, with a focus on sound source localization, speech enhancement, and speech separation.

Xiaofei Li received the Ph.D. degree in electronics from Peking University, Beijing, China, in 2013. He is currently a Postdoctoral Researcher with INRIA (French Computer Science Research Institute), Montbonnot-Saint-Martin, France. His research interests include multimicrophone speech processing for sound source localization, separation and dereverberation, single microphone signal processing for noise estimation, voice activity detection, and speech enhancement.