Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 6783184, 11 pages
https://doi.org/10.1155/2022/6783184

Research Article

A Multiple Sound Source Localization and Counting Method Based on the Kernel Density Estimator

Yuzhuo Fang,[1] **Xian Zang**,[1] **Juan Yang**,[1] **Hongcheng Zhou**,[1] **and Zhiyong Xu**[2]

[1]*School of Electronics and Information Engineering, Jinling Institute of Technology, Nanjing, Jiangsu, China*
[2]*School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, China*

Correspondence should be addressed to Yuzhuo Fang; fangjit@jit.edu.cn

Received 21 August 2021; Revised 19 December 2021; Accepted 10 January 2022; Published 29 January 2022

Academic Editor: Jelena Nikolić

An unambiguous signal processing algorithm when using a wide intermicrophone distance is proposed in this paper for simultaneously locating and counting multiple active sound sources. Based on the kernel density estimator, a multistage structure in the time-frequency domain is used to suppress the influence of spatial aliasing, then the pooled angular spectrum is combined with a peak search method having an updated cut-off threshold and a source merging module. Complete source localization and counting is realized through the combination of these two steps. Simulation results show that the proposed method has a more robust performance than the classic counterpart, especially in adverse environments with spatial aliasing, reverberation, and interference between different sound sources.

1. Introduction

Multiple sound source localization is a key element in many applications ranging from national security to video conference [1, 2]. Especially in far-field speech interaction, localizing multiple active sound sources accurately is an essential prerequisite for high-quality speech enhancement and recognition [3, 4]. In these applications, there is often ambient noise, reverberation, and mutual interference between different sources, which makes the source localization performance seriously affected.

In recent years, based on W-disjoint orthogonality (WDO) [5, 6] of observed signals between pairwise microphones, multiple sound source localization methods can be further divided into clustering, histogram, and angular spectrum.

The clustering method, such as interphase difference (IPD) [7] and direction estimation of mixing matrix (DEMIX) [8], directly achieves localization results by iteratively clustering time-frequency (TF) bins associated with each sound source [9]. It is sensitive to the initial clustering parameter [10]. The histogram method computes the weighted histogram such as circular integrated cross spectrum (CICS) [11] and smooth histogram [12] to make it have high estimation accuracy. However, the used array topology is relatively difficult to implement and popularize. Most of the above two methods do not consider spatial aliasing [13] which may exist in multiple sound source localization.

From the space sampling theorem, the intermicrophone distance should be smaller than half the minimum signal wavelength (e.g. 2 cm for a sampling rate of 16 kHz). When the distance exceeds half the wavelength, it can be regarded as a wide distance, and spatial aliasing generated from the high-frequency part will occur. The wider the distance, the more serious the aliasing. The excessively wide distance will have a significant impact on sound source localization. To obtain a high resolution in the time difference of arrival (TDOA) space, the arrangement of the microphones often uses a distance far exceeding half the wavelength, especially in the case of a small number of microphones, thus inevitably generating the influence of aliasing [14]. Hence, we focus on the angular spectrum method because of its applicability for a wide intermicrophone distance [10, 15]. It consists of two steps: (i) spectrum construction and (ii) localization and counting.

In the first step, the angular spectrum is constructed by accumulating the local function related to all possible angles in each TF bin. Classical generalized crosscorrelation (GCC) [16] computes the crosspower spectrum phase [17] from the pairwise microphones. Because of the limitations of the ideal single-source propagation model [18], the GCC spectrum will be severely distorted in environments with the coexistence of reverberation and spatial aliasing. Generated from the generalized state coherence transform (GSCT) [19, 20], approximate kernel density estimator (KDE) using the nonlinear Gaussian kernel [20, 21] has a significantly improved angular spectrum than GCC under reverberant environment. The frequency-dependent weighting factor of the Gaussian kernel can suppress spatial aliasing to a certain extent, but the effect is limited [14].

Existing sound source localization methods often regard the number of sound sources as a prior information [7, 22]. However, during the utterance of mixed speech signal, some sources may only last for a short period of time and the number of sources often changes dynamically [23], which makes it difficult to determine the number of sources in advance.

Therefore, the second step requires the active sources to be located and counted simultaneously [8] from the constructed spectrum. Current source localization and counting methods mainly use iterative search methods, such as single-point peak amplitude in peak search (PS) [24, 25], inner product in matching pursuit (MP) [26, 27], and source contribution rate in iterative contribution removal (ICR) [15]. Each iteration selects the optimal value satisfying the corresponding conditions, removes the components corresponding to the current sound source from the spectrum, and restarts the next. MP and ICR are more accurate than PS when obvious distortion is not produced in the spectrum. However, in a high-resolution spectrum, the computation cost of the inner product and the contribution rate is much higher, which is not conductive to real-time tracking the number of sources at different times. When the spectrum is seriously distorted due to reverberation and spatial aliasing, the current incorrect estimation will deteriorate the reconstructed spectrum, which may exert considerable influence on the following iteration [28].

In this paper, to effectively solve the spatial aliasing when using a wide intermicrophone distance, an unambiguous angular spectrum-based algorithm is proposed for simultaneously localizing and counting multiple active sound sources. In the spectrum construction step, the local KDE spectrum in each TF bin between the pairwise microphones is generated from a Gaussian kernel function; then a multistage structure [14] is used to divide the entire band into several sub-bands by the maximum unambiguous frequency. The sub-band spectra are pooled together across all the TF bins to suppress the influence of spatial aliasing. In the localization and counting step, PS is used considering the simplicity and real-time processing. An updated cut-off threshold according to the current peak amplitude is used to replace the fixed threshold in traditional PS, which strengthen the flexibility of counting. Then the preliminary estimation is passed through a source merging module [15] to eliminate the duplicate sound source.

The remainder of this paper is organized as follows: in Section 2, the KDE spectrum is constructed on the basis of the signal propagation model of multiple sound sources. In Section 3, the MS structure is used to construct the KDEMS spectrum. Then PS are combined to form the complete localization and counting process. Numerical comparisons between the proposed KDEMS-PS and the other classical ones are given in Section 4. Finally, Section 5 concludes the paper.

2. Signal Propagation Model of Multiple Sound Sources and KDE Spectrum Construction

Set $\{s_n(i) \mid n = 1, 2, \ldots, N\}$ as N sound source signals at the transmitting end, and $\{x_m(i) \mid m = 1, 2, \ldots, M\}$ as the observed signals corresponding to M microphones at the receiving end. The sampling rate is f_s. Then in the approximately far field, the discrete time signal propagation model of multiple sound sources can be expressed as follows:

$$x_m(i) = \sum_{n=1}^{N} \mathbf{h}_{m,n}^T s_n(i) + v_m(i), \tag{1}$$

where $h_{m,n} = [\mathbf{h}_{m,n}(0), \ldots, h_{m,n}(L-1)]^T$ denotes the vector of length L corresponding to the impulse response between the n-th source and the m-th microphone, $\mathbf{s}_n(i) = [s_n(i), \ldots, s_n(i-L+1)]^T$ denotes the vector corresponding to the n-th source, and $v_m(i)$ denotes the additive white Gaussian noise of the m-th microphone which is independent of source signals and impulse responses.

Through N_{FFT} point short-time Fourier transform (STFT), the expression in the discrete TF domain can be obtained as

$$X(r,k) = \sum_{n=1}^{N} \mathbf{H}_n(k) S_n(r,k) + \mathbf{V}(r,k), \tag{2}$$

where r and k denote the frame and frequency indices, respectively, $\mathbf{X}(r,k) = [X_1(r,k), \ldots, X_m(r,k), \ldots, X_M(r,k)]^T$, $S_n(r,k)$ and $X_m(r,k)$ denote the STFT coefficients corresponding to $s_n(i)$ and $x_m(i)$, respectively, $\mathbf{V}(r,k) \in \mathbb{C}^{M\times 1}$ denotes the complex vector corresponding to the noise, and

$$\mathbf{H}_n(k) = \left[H_{1,n}(k), \ldots, H_{m,n}(k), \ldots, H_{M,n}(k)\right]^T, \tag{3}$$

where $H_{m,n}(k)$ is the transfer function between the n-th source and the m-th microphone, including the direct wave component $H_{m,n}^{(\mathrm{dir})}(k)$ and the reverberation component $H_{m,n}^{(\mathrm{rev})}(k)$. Since the impulse response is time-invariant, the transfer function is only related to k.

When the intensity of the reflected wave is uniformly distributed in all possible directions of propagation, $H_{m,n}^{(\mathrm{rev})}(k)$ can be regarded as spatially diffuse noise [29, 30]. $H_{m,n}(k)$ can be decomposed as

$$H_{m,n}(k) = H_{m,n}^{(\mathrm{dir})}(k) + H_{m,n}^{(\mathrm{rev})}(k), \tag{4}$$

where $H_{m,n}^{(\mathrm{dir})}(k) = \alpha_{m,n}\exp(-j2\pi f_k i_{m,n}/f_s)$, f_k represents the frequency at the k-th frequency bin, $\alpha_{m,n}$ and $(i_{m,n}/f_s)$ denote the corresponding propagation attenuation and time, respectively. Under anechoic, noise-free, and WDO conditions [5], only $H_{m,n}^{(\mathrm{dir})}(k)$ exists, and at most one sound source energy dominates each TF bin. Then equation (2) can be simplified as follows:

$$\mathbf{X}(r,k) = \begin{bmatrix} \alpha_{1,\eta(r,k)}(k)\exp\left(\dfrac{-j2\pi f_k i_{1,\eta(r,k)}}{f_s}\right) \\ \vdots \\ \alpha_{M,\eta(r,k)}(k)\exp\left(\dfrac{-j2\pi f_k i_{M,\eta(r,k)}}{f_s}\right) \end{bmatrix} S_{\eta(r,k)}(r,k), \tag{5}$$

where $\eta(r,k) \in \{1,\ldots,N\}$ is the index of the dominant sound source in the current TF bin.

Considering one pair of microphones m_a, m_b in the array, the normalized cross-power spectrum (NCS) of the observed signal can be expressed as [21].

$$\mathrm{NCS}(r,k) = \frac{X_a(r,k)X_b^*(r,k)}{\left|X_a(r,k)X_b^*(r,k)\right|} = \exp\left(-j2\pi f_k \tau_{\eta(r,k)}\right), \tag{6}$$

where $(\tau_{\eta(r,k)} = (i_{a,\eta(r,k)} - i_{b,\eta(r,k)})/f_s)$ denotes the true TDOA of the dominant sound source between m_a and m_b. Without loss of generality, the subscript of $\tau_{\eta(r,k)}$ can be omitted and τ can be estimated by regarding it as a random variable that satisfies a given probability distribution. Then a Gaussian kernel function is introduced as [20].

$$g(e(\tau)) = \frac{1}{2\pi f_k}\exp\left(-\frac{(e(\tau)/2\pi f_k)^2}{2h_K^2}\right), \tag{7}$$

where

$$e(\tau) = \left|\exp(-j2\pi f_k\tau) - \mathrm{NCS}(r,k)\right|, \tag{8}$$

is a function of τ to calculate the Euclidean distance between $\exp(-j2\pi f_k\tau)$ and $\mathrm{NCS}(r,k)$,

$$h_K = \frac{\tau_{\max}}{B} = \frac{d_{\max}}{cB}, \tag{9}$$

is the bandwidth of the kernel function where $\tau_{\max}$ is the maximum possible TDOA, $d_{\max}$ is the maximum spacing between adjacent microphones across the given pairs in the array, c is the velocity of sound propagation, and B is a factor that affects the resolution in the TDOA space. If B is set too large, the kernel function may generate many burrs, which may subsequently lead to the distortion of the spectrum. If B is set too small, the main lobe is obese, which is not conducive to locate the source correctly. B is set as 20 empirically when f_s = 16kHz. Substitute equation (8) into equation (7), the local KDE spectrum in each TF bin can be expressed as

$$\varphi(r,k,\tau) = \frac{1}{2\pi f_k}\exp\left(-\frac{\left|\exp(-j2\pi f_k\tau) - \mathrm{NCS}(r,k)\right|^2}{2\sigma_k^2}\right), \tag{10}$$

where $\sigma_k = 2\pi f_k h_K$.

Pool the local spectra across all the TF bins [10, 21]; the total KDE spectrum can be constructed as

$$\Phi_{\mathrm{KDE}}(\tau) = \sum_k \varphi_k(\tau) = \sum_{(r,k)} \varphi(r,k,\tau), \tag{11}$$

where $\varphi_k(\tau)$ denotes the narrow-band KDE spectrum corresponding to the k-th frequency bin after pooling the local spectra across all time frames. The estimated TDOAs of sound sources can be obtained from the peaks of $\Phi_{\mathrm{KDE}}(\tau)$.

3. Localization and Counting Based on Multistage Structure and Peak Search

There is a frequency-dependent weighting factor $(1/(2\pi f_k))$ in $\varphi(r,k,\tau)$, which means that the higher the frequency is, the more the amplitude of the local spectrum is suppressed. Then spatial aliasing mainly caused by the high-frequency part is weakened to a certain degree, but it cannot be completely eliminated [14], especially in a strong reverberation environment where the spectrum is severely distorted. Hence, a multistage structure is used to eliminate the influence of spatial aliasing more efficiently in this section.

3.1. Spectrum Construction with Multistage Structure. When calculating $\Phi_{\mathrm{KDE}}(\tau)$ in equation (11), the indices of the lowest frequency bin and the highest are set to k_L and k_H, respectively. Then the corresponding center frequencies can be set to $(f_L = k_L \cdot f_s/N_{\mathrm{FFT}})$ and $(f_H = k_H \cdot f_s/N_{\mathrm{FFT}})$, respectively.

According to the spatial Nyquist sampling theorem, the unambiguous condition can be expressed as

$$d_{\mathrm{mic}} \le \frac{\lambda_{\min}}{2}, \tag{12}$$

where d_{mic} is the distance between m_a and m_b, and $\lambda_{\min}$ represents the minimum signal wavelength corresponding to f_H which can be expressed as

$$\lambda_{\min} = \frac{c}{f_H}. \tag{13}$$

Substituting equation (13) into equation (12), the condition can be rewritten as

$$f_H \le \frac{c}{2d_{\mathrm{mic}}} = f_{\mathrm{UA}}, \tag{14}$$

where f_{UA} denotes the maximum unambiguous frequency. If equation (14) is fulfilled, no spatial aliasing exists. However, in order to produce a higher resolution angular spectrum to distinguish different sound sources, it is necessary to use a wider d_{mic}, which makes f_H exceed f_{UA} inevitably.

If $f_H > f_{\mathrm{UA}}$, the entire frequency band can be divided into two sub-bands $[f_L, f_{\mathrm{UA}}]$ and $[f_{\mathrm{UA}}, f_H]$ where a single sub-band may contain one or more continuous frequency bins. Then if $f_H > 2f_{\mathrm{UA}}$, $[f_{\mathrm{UA}}, f_H]$ can be further divided into $[f_{\mathrm{UA}}, 2f_{\mathrm{UA}}]$ and $[2f_{\mathrm{UA}}, f_H]$. By analogy, if $f_H > pf_{\mathrm{UA}}$ where $p > 1$, the sub-bands divided at the p-th stage can be deduced as $[(p-1)f_{\mathrm{UA}}, pf_{\mathrm{UA}}]$ and $[pf_{\mathrm{UA}}, f_H]$. The entire

band can be finally divided into a total of P sub-bands $[f_L, f_{\mathrm{UA}}], \ldots, [(P-1)f_{\mathrm{UA}}, f_H]$. P can be obtained as

$$P = \mathrm{ceil}\left(\frac{f_H - f_L}{f_{\mathrm{UA}}}\right), \tag{15}$$

where $\mathrm{ceil}(\cdot)$ denotes the round-up operator.

Then by pooling $\varphi_k(\tau)$ across all frequency bins contained in each sub-band, the sub-band KDE spectrum can be obtained as

$$\Phi_{\mathrm{KDE}}^{(p)}(\tau) = \sum_{k=k_{pL}}^{k_{pH}} \varphi_k(\tau), \tag{16}$$

where $p = 1, \ldots, P$, k_{pL} denotes the index of the lowest frequency bin contained in the p-th sub-band which can be expressed as

$$k_{pL} = \begin{cases} k_L, & p = 1, \\ (p-1)k_{\mathrm{UA}} + 1, & p = 2, \ldots, P, \end{cases} \tag{17}$$

and k_{pH} denotes the index of the highest frequency bin which can be expressed as

$$k_{pH} = \begin{cases} pk_{\mathrm{UA}}, & p = 1, \ldots, P-1, \\ k_H, & p = P. \end{cases} \tag{18}$$

In equations (17) and (18), k_{UA} is the index corresponding to f_{UA} which can be obtained as

$$k_{\mathrm{UA}} = \mathrm{floor}\left(\frac{f_{\mathrm{UA}} \cdot N_{\mathrm{FFT}}}{f_s}\right), \tag{19}$$

where $\mathrm{floor}(\cdot)$ is the round-down operator. When $p = 1$, the frequency will not exceed f_{UA}, so there is no interference of spatial aliasing; when $p > 1$, equation (14) is no longer fulfilled, and the p-th sub-band spectrum $\Phi_{\mathrm{KDE}}^{(p)}(\tau)$ will have at most one more false peak than the previous sub-band.

Then the MS structure is used for sub-band processing, and the output of the first stage can be expressed as

$$\Phi_{\mathrm{KDEMS}}^{(1)}(\tau) = \Phi_{\mathrm{KDE}}^{(1)}(\tau). \tag{20}$$

By multiplying $\Phi_{\mathrm{KDEMS}}^{(1)}(\tau)$ with the second sub-band spectrum, the output of the second stage can be obtained as

$$\Phi_{\mathrm{KDEMS}}^{(2)}(\tau) = \Phi_{\mathrm{KDEMS}}^{(1)}(\tau)\Phi_{\mathrm{KDE}}^{(2)}(\tau) = \Phi_{\mathrm{KDE}}^{(1)}(\tau)\Phi_{\mathrm{KDE}}^{(2)}(\tau), \tag{21}$$

where the false peaks in $\Phi_{\mathrm{KDE}}^{(2)}(\tau)$ can be suppressed by $\Phi_{\mathrm{KDEMS}}^{(1)}(\tau)$. Then the weighted output $\Phi_{\mathrm{KDEMS}}^{(2)}(\tau)$ is substituted into the next stage. By analogy, the output of the P-th stage, that is, the KDEMS spectrum can be deduced as

$$\Phi_{\mathrm{KDEMS}}(\tau) = \Phi_{\mathrm{KDEMS}}^{(P)}(\tau) = \Phi_{\mathrm{KDEMS}}^{(P-1)}(\tau)\Phi_{\mathrm{KDE}}^{(P)}(\tau) = \prod_{p=1}^{P} \Phi_{\mathrm{KDE}}^{(p)}(\tau). \tag{22}$$

The spectrum of the lower sub-band has a wider main lobe and lower resolution but is less affected by spatial aliasing, while the spectrum of the higher sub-band has a narrower main lobe and higher resolution but is subject to spatial aliasing. Combining these two parts of the spectrum through the MS structure can effectively absorb the advantages of each part while making up for their shortcomings, resulting in the final output spectrum with obvious ambiguity suppression and high resolution.

Comparing equations (11) and (22), it can be seen that the MS structure does not increase the computational complexity. The total number of operations for both KDE and KDEMS can be approximately expressed as

$$N_{\mathrm{com}} = N_\varphi \times N_\tau \times (k_H - k_L + 1), \tag{23}$$

where N_φ denotes the number of operations required for $\varphi_k(\tau)$, and N_τ denotes the number of samples contained in the TDOA space.

3.2. Multiple Sound Source Localization and Counting by Peak Search. When the angular spectrum is constructed, an improved PS is used to realize multiple sound source localization and counting. Without loss of generality and to facilitate the subsequent processing, the normalized form of the spectrum is obtained as

$$\Phi(\tau) = \frac{\Phi_{\mathrm{AL}}(\tau) - \min(\Phi_{\mathrm{AL}}(\tau))}{\max(\Phi_{\mathrm{AL}}(\tau)) - \min(\Phi_{\mathrm{AL}}(\tau))}, \tag{24}$$

where AL represents KDE, KDEMS, or other angular spectrum-based methods (e.g. GCC) and $\min(\cdot)$ and $\max(\cdot)$ represent the operators to find the minimum and maximum values in the function, respectively. Then the sequence containing all the peaks in $\Phi(\tau)$ can be expressed as

$$\boldsymbol{\Phi} = \left[\Phi(\tau_1), \ldots, \Phi(\tau_Q)\right], \tag{25}$$

where $\tau_1, \ldots, \tau_Q$ denote the TDOAs corresponding to the peaks in the TDOA space sampled with τ_{grid} as the grid value and Q denotes the total number of peaks. The elements in equation (25) are sorted in the descending order of peak amplitude, so the condition $\Phi(\tau_i) \geq \Phi(\tau_{i+1})$ where $i = 1, \ldots, Q-1$ is fulfilled. When the number of sound sources N is known where $N \leq Q$, only the first N elements need to be extracted from Φ, and the set of the estimated TDOAs can be expressed as $\{\tau_i | i = 1, \ldots, N\}$. However, in most real cases, N is unknown. Hence, it is necessary to count the number of sound sources to obtain effective localization results.

In traditional PS, a fixed cut-off threshold set as Γ is compared with the elements in Φ one by one iteratively according to the order of the sequence until the element is not greater than Γ. If Γ is set too large or too small, it will lead to excessive missing alarm rate or false alarm rate, respectively. So the threshold is set to vary as the peak amplitude changes to improve the counting flexibility. Set Γ_1 as the first threshold. It is generally assumed that there is at least one active sound source, so $\Phi(\tau_1) > \Gamma_1$ holds. Then the information of the first peak amplitude is introduced into the threshold setting and the threshold used in the next iteration can be updated as $\Gamma_2 = \max\{\Phi(\tau_1)/2, \Gamma_1\}$ where $\max\{\cdot\}$ represents the operators to find the maximum in the set. If

$\Phi(\tau_2) > \Gamma_2$, the iteration continues. Set the index of the iteration to i. Then the threshold used in the i-th iteration can be updated as

$$\Gamma_i = \max\{\Phi(\tau_{i-1})/2, \Gamma_1\}, \tag{26}$$

where $i = 2, \ldots, I$. I is the total number of iterations. Give the following condition:

$$\begin{cases} \Phi(\tau_i) \le \Gamma_i \\ i > I_{\max} \end{cases}, \tag{27}$$

where $I_{\max}$ denotes the maximum number of iterations. When either of the two conditions in equation (27) is not fulfilled, the iteration stops. The set of the estimated TDOAs can be expressed as

$$\Omega = \{\tau_i | i = 1, \ldots, I-1\}. \tag{28}$$

Due to the interference of reverberation and noise, there may be multiple peaks with similar amplitude near the peak corresponding to the true TDOA in the distorted angular spectrum, which will lead to repeated estimation of the same sound source, thus bringing undesirable counting results. To address this problem, it is necessary to merge the duplicate estimation. The source merging is implemented as follows: for any two estimated TDOAs τ_i and $\tau_{i'}$ where $i \neq i'$, the condition can be given as

$$|\tau_i - \tau_{i'}| < \tau_{\min}, \tag{29}$$

where $\tau_{\min}$ indicates the allowed minimum distance between the estimated TDOAs corresponding to two different sound sources. $\tau_{\min}$ can be obtained as [15]

$$\tau_{\min} = \frac{d_{\mathrm{mic}} \sin(A_{\min})}{c}, \tag{30}$$

where $A_{\min}$ is the minimum angular distance empirically set to 10° when the interval is $[0, 180]^\circ$. When equation (29) is fulfilled, according to the peak amplitude, the estimated results can be reassigned as

$$\begin{aligned} &\tau_{i'} = \tau_i, \quad \Phi(\tau_i) \ge \Phi(\tau_{i'}), \\ &\tau_i = \tau_{i'}, \quad \text{otherwise.} \end{aligned} \tag{31}$$

After the reassignment, all the estimated TDOAs with the same value are merged into only one. Then the final localization and counting results can be expressed as

$$\Omega' = \{\widehat{\tau}_n | n = 1, \ldots, \widehat{N}\}, \tag{32}$$

where $\Omega' \subseteq \Omega$, $\widehat{\tau}_n$ indicates the estimated TDOA of the n-th sound source when using PS with the source merging module, and $\widehat{N} = \mathrm{card}(\Omega')$ indicates the estimated number of sound sources where $\mathrm{card}(\cdot)$ is the operator to find the number of elements in the set.

A complete localization and counting method can be used by combining the constructed angular spectrum with PS, where the combination is marked with "-" in this paper. Hence, when the angular spectrum is generated by GCC, KDE, or KDEMS, the corresponding localization and counting method can be called GCC-PS, KDE-PS or KDEMS-PS, respectively. The block diagram of the proposed KDEMS-PS is shown in Figure 1, and the steps can be summarized as follows:

(i) STFT: Transform the observed signals of one pair of microphones $x_a(i)$, $x_b(i)$ into the corresponding coefficients $X_a(r,k)$, $X_b(r,k)$ in each TF bin by STFT.

(ii) Narrowband KDE spectrum calculation: Calculate $\mathrm{NCS}(r,k)$ from equation (6), and then pool the local KDE spectrum $\varphi(r,k,\tau)$ in equation (10) across the k-th frequency bin to obtain the narrowband KDE spectrum $\varphi_k(\tau)$.

(iii) Multi-stage processing: Determine the number of sub-bands P and the index corresponding to the maximum unambiguous frequency k_{UA} from equations (15) and (19), respectively, then obtain the sub-band KDE spectrum $\Phi_{\mathrm{KDE}}^{(p)}(\tau)$ where $p = 1, \ldots, P$ from equation (16).

(iv) Pooling and normalization: Pool all the sub-band spectra to construct the KDEMS angular spectrum $\Phi_{\mathrm{KDEMS}}(\tau)$ from equation (22), then obtain the normalized form $\Phi(\tau)$ from equation (24).

(v) Peak search: Search the peaks in $\Phi(\tau)$, sort them in descending order, and then compare them with the updated cut-off threshold Γ_i from updated equation (26) to obtain the estimated TDOAs in Ω.

(vi) Source merging: Merge the repeatedly estimated sound sources from equation (31) to obtain the final localization and counting results in Ω'.

4. Numerical Analysis

In this section, compared with classic GCC and KDE, the KDEMS spectrum is analyzed. Then the localization and counting performance of the combined GCC-PS, KDE-PS, and KDEMS-PS is further discussed.

The sound sources are taken from 16 pure speeches composed of 8 male and 8 female in the TIMIT database [31]. Each speech segment sampled with $f_s = 16\mathrm{kHz}$ lasts 2 seconds, and has the same average power through preprocessing. The room is $7.8\mathrm{m} \times 8.1\mathrm{m} \times 3\mathrm{m}$ whose x-y plane diagram is shown in Figure 2. A pair of omni-directional microphones m_a, m_b with distance d_{mic} parallel to the x-axis is located at the center of the room marked with o, where d_{mic} is set to be an multiple of $0.5\lambda_{\min}$ with a positive integer u as the multiple factor. N sound sources are distributed on a semicircle with o as the centroid, where direction of arrival (DOA) of the n-th sound source $\theta_n \in [0, 180]^\circ$ is defined in an anti-clockwise manner with 90° being the direction perpendicular to the line connecting m_a and m_b. Then, the true TDOA can be obtained as $(\tau_n = -d_{\mathrm{mic}} \cos(\theta_n)/c)$ where c is the sound velocity set as $(344m/s)$. d_{ms} is the microphone-source distance set as 3m. In Figure 2, z coordinates of the sound sources and the microphones are all set as 1.3m.

When the reverberation time expressed as RT_{60} changes between $200ms$ and $500ms$ and d_{mic} changes from $0.5\lambda_{\min}$ to $6\lambda_{\min}$ with $0.5\lambda_{\min}$ as the interval, 200 simulations are

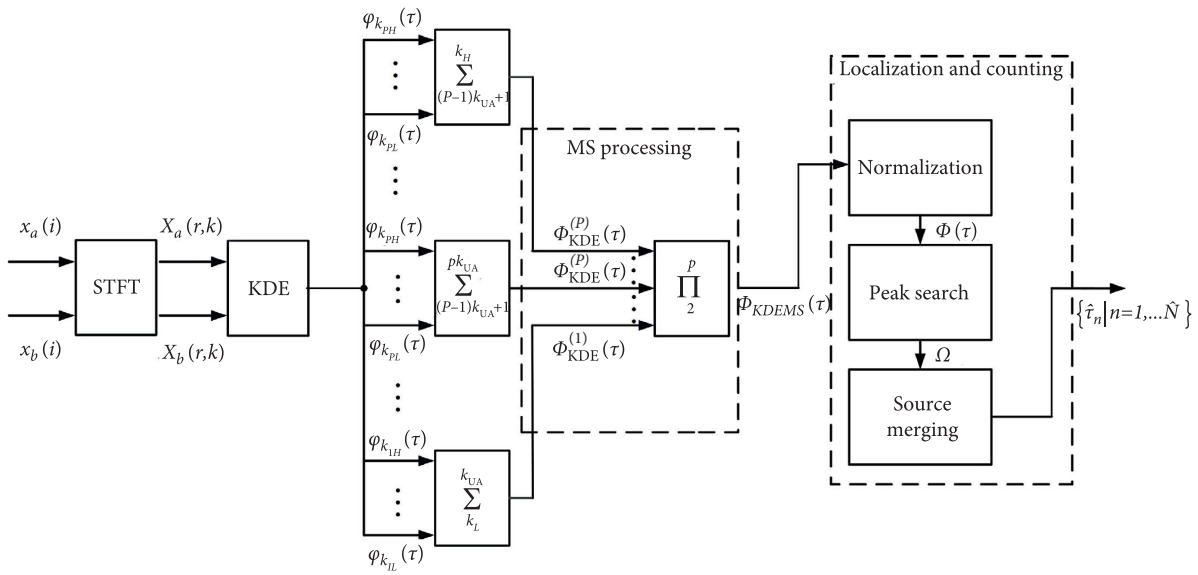

Figure 1: A block diagram of the multiple sound source localization and counting method combined KDEMS with PS.

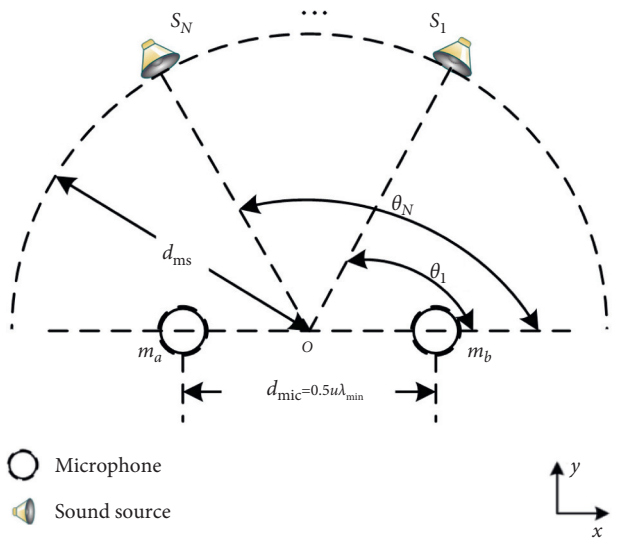

Figure 2: The x-y plane diagram of the room.

performed under each scenario. In each simulation, N sound sources are randomly selected from the 16 candidate segments, convolved with the room impulse response (RIR) generated by the image-source model [32, 33], and added white Gaussian noise with fixed signal-to-noise ratio (SNR) which can be obtained as

$$\mathrm{SNR} = 10\log_{10}\frac{P_a + P_b}{2P_v}, \tag{33}$$

where P_a and P_b indicate the average power of the noise-free observed signal corresponding to m_a and m_b, respectively, and P_v indicates the average power of the additive white Gaussian noise at the receiving end. In the simulation scenarios, multiple sound sources with equal angular intervals are used. When N is 2,4, or 6, the true DOA distribution is shown in Table 1.

Table 1: Number of sources N versus DOA.

N	DOA (°)
2	60, 120
4	30, 60, 120, 150
6	0, 30, 60, 120, 150, 180

Then the observed signal is passed through the bandpass filters between 0.2 kHz and 4 kHz [14] and transformed into the T-F coefficients with a 1024-point Hamming-weighted STFT, where the frame shift rate is 25%. Thirty consecutive frames are extracted to pool the spectrum. According to the current environment, the parameters are chosen empirically. In the Gaussian kernel function used by KDE and KDEMS, B is set to 20. τ_{grid} in the TDOA space is uniformly set to 1×10^{-5} s. In the localization and counting step, Γ_1 is set to 0.35, and $I_{\max}$ is set to 10.

When $N = 2$ and $RT_{60} = 200$ ms, with d_{mic} set as $0.5\lambda_{\min}$, $1.5\lambda_{\min}$, $3.5\lambda_{\min}$, or $5.5\lambda_{\min}$, the normalized spectra generated by GCC, KDE, and KDEMS are shown in Figure 3. In each subfigure, three curves colored by red, green, and blue correspond to GCC, KDE, and KDEMS, respectively. For the intuitiveness and unity of graphic expression, a weighting factor $(1/\tau_{\max})$ is introduced to ensure the unified abscissa with different d_{mic}. Then the true TDOAs of the sound sources are located at 0.5 and −0.5, which are marked with two vertical dashed lines.

In Figure 3(a), since equation (12) is fulfilled when $d_{\mathrm{mic}} = 0.5\lambda_{\min}$, there is no spatial aliasing. Hence, the KDEMS spectrum coincides with KDE. However, the low-resolution spectra generated by the narrow d_{mic} makes the peak too wide to produce good discrimination between different sound sources in the TDOA space. GCC has only one peak obviously deviated from the true TDOAs which make it difficult to give correct estimation.

In Figure 3(b), when d_{mic} is widened to $1.5\lambda_{\min}$, the resolution in the TDOA space improves and all the three

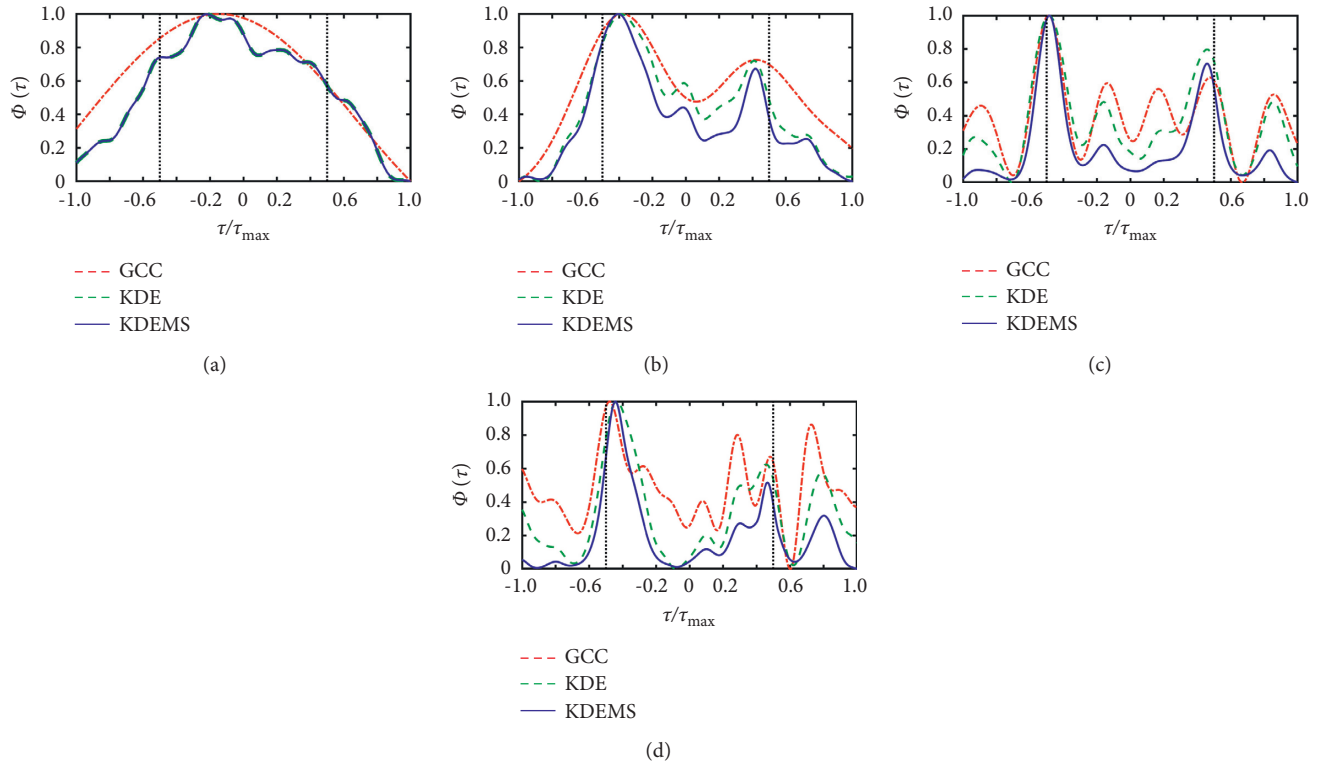

FIGURE 3: Normalized spectral function $\Phi(\tau)$ of GCC, KDE, and KDEMS versus $(\tau/\tau_{\max})$ when $RT_{60} = 200$ ms and (a) $d_{\text{mic}} = 0.5\lambda_{\min}$; (b) $d_{\text{mic}} = 1.5\lambda_{\min}$; (c) $d_{\text{mic}} = 3.5\lambda_{\min}$; (d) $d_{\text{mic}} = 5.5\lambda_{\min}$.

spectra can form two peaks with the two highest amplitude near the true TDOAs. However, the main lobe corresponding to each sound source is still too wide and there is still a certain degree of deviation.

In Figure 3(c), when d_{mic} is further widened to $3.5\lambda_{\min}$, sharper peaks close to the true TDOAs make the deviation smaller than Figure 3(b). However, the simultaneously introduced spatial aliasing makes false peaks with a certain amplitude start to appear at −0.9, −0.2, 0.2, and 0.9 in GCC. Due to the suppression of high frequency components by the frequency-dependent $(1/(2\pi f_k))$ in Gaussian kernel function, the amplitudes of the false peaks in KDE are lower than GCC. Then by further MS processing, the false peaks are almost completely suppressed in KDEMS.

In Figure 3(d), when $d_{\text{mic}} = 5.5\lambda_{\min}$, the number of false peaks in GCC is obviously more than KDE and KDEMS, the overall amplitudes of the false peaks becomes higher than Figure 3(c). The amplitude at −0.3 in GCC is almost equivalent to the one at 0.5 corresponding to the true TDOA, which is easy to produce incorrect estimation. However, KDE and KDEMS maintain sharp peaks near the true TDOAs and still have a good suppression on the false peaks, thus have better recognition than GCC.

The normalized spectra when RT_{60} increases to 500ms are shown in Figure 4. The enhancement of reverberation makes the spectrum distortion more serious. In Figure 4(a), GCC, KDE, and KDEMS still face the influence of low resolution. In Figure 4(b), all the three spectra cannot estimate the TDOAs effectively because of the deviation. Compared with Figure 3(c), the amplitudes of the false peaks at −0.9, −0.2, 0.2, and 0.9 are higher in Figure 4(c), where the amplitude at −0.2, 0.2 is almost equivalent to the one at 0.5 in GCC. In Figure 4(d), the amplitudes of the false peaks at 0.3, 0.7 have exceeded the amplitude at 0.5 in GCC, the amplitude at 0.7 is close to 0.5 in KDE, while KDEMS has a relatively flat spectrum which shows that KDEMS is more robust than GCC and KDE under strong reverberation.

After the spectrum analysis of a single simulation, the source localization and counting performance of GCC-PS, KDE-PS, and KDEMS-PS can be evaluated in terms of three measures defined as

$$R_{\text{rate}} = \frac{\widehat{N}_{\text{co}}}{N}, P_{\text{rate}} = \frac{\widehat{N}_{\text{co}}}{\widehat{N}}, F_{\text{score}} = \frac{2R_{\text{rate}} \cdot P_{\text{rate}}}{R_{\text{rate}} + P_{\text{rate}}}, \tag{34}$$

where recall rate expressed as R_{rate} and precision rate expressed as P_{rate} are used to measure miss-detections and false alarms, respectively, F-score expressed as F_{score} is the combination of R_{rate} and P_{rate}, and $\widehat{N}_{\text{co}}$ denotes the number of correctly estimated sound sources. The condition for correct estimation can be expressed as

$$|\widehat{\theta} - \theta| < 5^{\circ}, \tag{35}$$

where $\widehat{\theta}$ and θ represent the DOAs corresponding to the estimated TDOA and the true TDOA, respectively.

Set d_{mic} as $3.5\lambda_{\min}$, the average values of these three measures versus SNR when $N = 2$ and $N = 6$ are shown in Figures 5 and 6, respectively, where they all reach its best

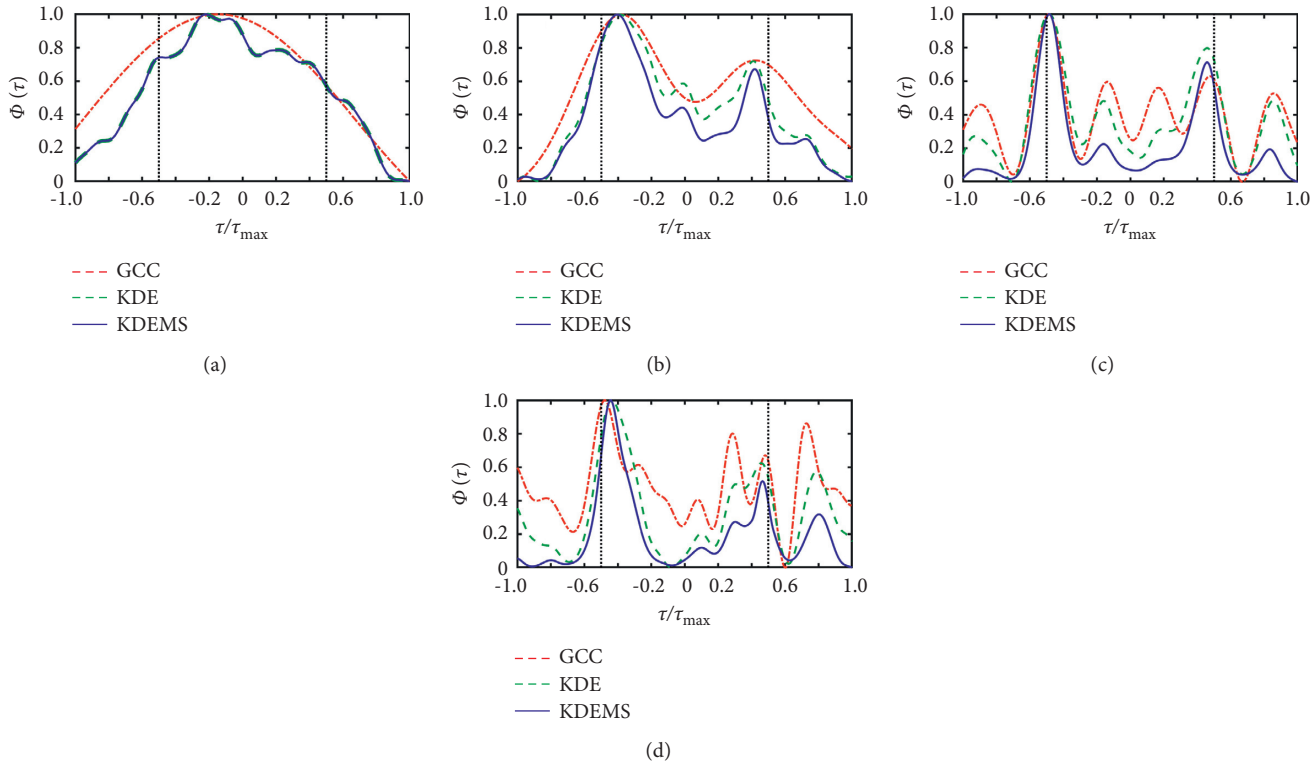

Figure 4: Normalized spectral function $\Phi(\tau)$ of GCC, KDE, and KDEMS versus (τ/τ_{max}) when $RT_{60} = 500$ ms and (a) $d_{mic} = 0.5\lambda_{min}$; (b) $d_{mic} = 1.5\lambda_{min}$; (c) $d_{mic} = 3.5\lambda_{min}$; (d) $d_{mic} = 5.5\lambda_{min}$.

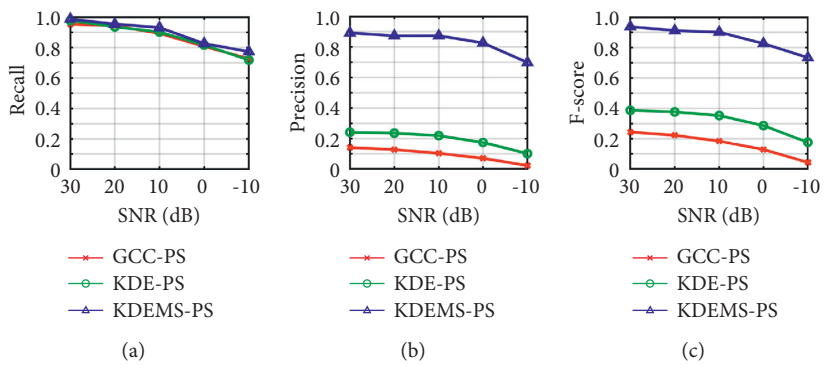

Figure 5: When $N = 2$ and $d_{mic} = 3.5\lambda_{min}$, the average measure of GCC, KDE and KDEMS versus SNR using (a) recall; (b) precision; (c) F-score.

value at 1 and worst at 0. As SNR decreases from 30 to −10 dB, the measures of GCC-PS, KDE-PS and KDEMS-PS gradually decline. In Figure 5(a), since only two sound sources are used, the difference of R_{rate} between these three methods is not very obvious. But the false peaks in GCC and KDE spectra will produce more false alarms than KDEMS. With the decrease of SNR, the increase of the spectrum base further aggravates the influence of false peaks, so P_{rate} of KDEMS-PS is significantly better than the other two methods. The situation of F_{score} is similar to that of P_{rate}.

Due to the increase in the number of sound sources, the missed-detection situation has been enhanced. Then R_{rate} of KDEMS-PS is significantly better than the other two in Figure 6(a). False alarms still exist in Figure 6(b) but will be alleviated to a certain extent by the added sources. P_{rate} of KDEMS-PS is still the best of the three methods. The situation of F_{score} is similar to that of R_{rate}.

It can be seen from Figures 5 and 6 that when the number of sound sources is small, P_{rate} can better distinguish the performance of different methods, and when the number of sound sources is large, R_{rate} is better, while F_{score} is the harmonic average of R_{rate} and P_{rate} to comprehensively evaluate the performance. Hence, without loss of generality, use F_{score} to measure the performance of different methods. With SNR set as 20 dB, the average F-score versus d_{mic} when $RT_{60} = 200$ ms and $RT_{60} = 500$ ms is shown in Figures 7 and 8, respectively.

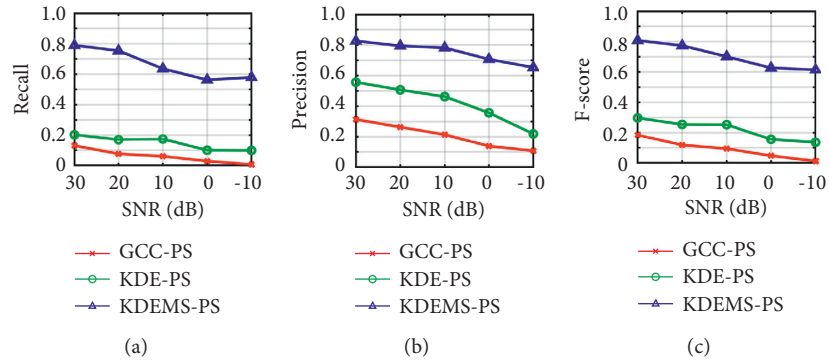

Figure 6: When $N = 6$ and $d_{\text{mic}} = 3.5\lambda_{\min}$, the average measure of GCC, KDE and KDEMS versus SNR using (a) recall; (b) precision; (c) F-score.

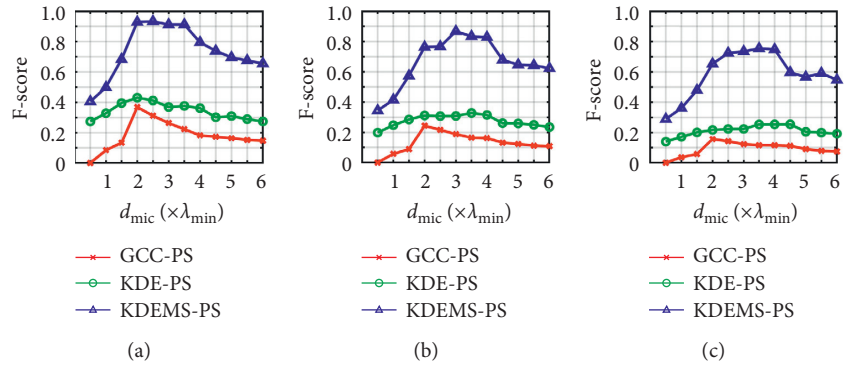

Figure 7: The average F-score of GCC-PS, KDE-PS, and KDEMS-PS versus d_{mic} when $RT_{60} = 200$ ms and (a) $N = 2$; (b) $N = 4$; (c) $N = 6$.

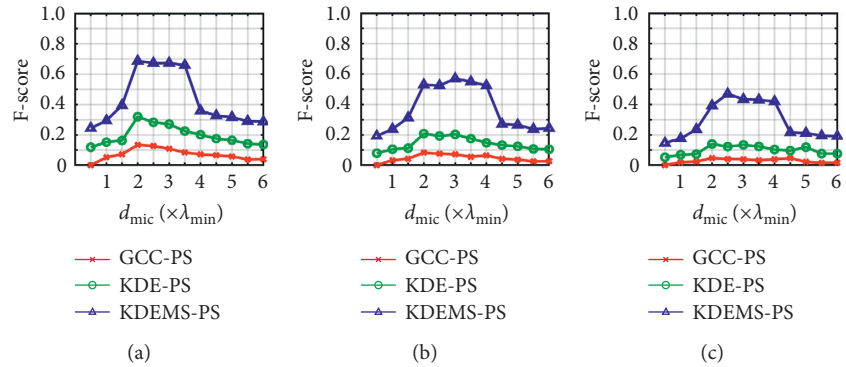

Figure 8: The average F-score of GCC-PS, KDE-PS, and KDEMS-PS versus d_{mic} when $RT_{60} = 500$ ms and (a) $N = 2$; (b) $N = 4$; (c) $N = 6$.

In Figure 7(a), when $d_{\text{mic}} \le 1.5\lambda_{\min}$, the deviation caused by the low resolution makes $\widehat{N}_{\text{co}}$ small, thus bringing low F_{score}, then F_{score} gradually rises as d_{mic} widens. When $d_{\text{mic}} > 1.5\lambda_{\min}$, the growth of F_{score} disappears, GCC-PS and KDE-PS begin to gradually decline; this is due to the increase in the amplitude of false peaks caused by spatial aliasing, which makes $\widehat{N}$ increase. However, due to MS processing, KDEMS-PS is much higher than the other two and it can stabilize around 0.9 with no obvious downtrend when $1.5\lambda_{\min} < d_{\text{mic}} \le 3.5\lambda_{\min}$, which demonstrates the capability to effectively suppress the influence of spatial aliasing.

The overall F_{score} declines to a certain extent when N increases. When $d_{\text{mic}} = 4\lambda_{\min}$, KDEMS-PS can maintain around 0.8 and 0.6 in Figures 7(b) and 7(c), respectively, not decline as shown in Figure 7(a). This is because the increase of N requires higher resolution to distinguish different sound sources, so a higher F_{score} can be brought about when d_{mic} is appropriately increased.

The rule in Figure 8 is similar to that of Figure 7. The enhancement of reverberation makes $\widehat{N}_{\text{co}}$ smaller on the one hand and $\widehat{N}$ larger on the other hand, thus makes F_{score} of different methods have a certain degree of decline. The

overall performance is still KDEMS-PS >$r$bin KDE-PS >$r$bin GCC-PS. When $1.5\lambda_{\min} < d_{\text{mic}} \leq 4\lambda_{\min}$, KDEMS-PS can still maintain above 0.5 and 0.4 when $N = 4$ and $N = 6$, respectively, much higher than KDE-PS and GCC-PS, which shows the robustness of the method under strong reverberation and interference between different sound sources.

5. Conclusion

To correctly locate the sound sources in real scene, the number of them needs to be estimated simultaneously. Based on the kernel density estimator, a multiple sound source localization and counting method called KDEMS-PS is proposed in this paper. Divide the entire frequency band into several sub-bands and process them stage by stage, the KDEMS angular spectrum is constructed. Then PS with an updated threshold and a source merging module is combined to locate and count multiple sound sources. As shown in the computer simulation using spectrum analysis and comparison of F-score, KDEMS spectrum can effectively weaken the interference caused by false peaks and KDEMS-PS is a robust multiple sound source localization and counting method with good spatial aliasing suppression when using a wide intermicrophone distance. In the experiment, we found that the MS structure is mainly used to process sound sources with more low-frequency energy and more uniform spectrum coverage (e.g. speech). When the spectral components are concentrated in the higher frequency range (e.g. bird sound), the sub-band processing may face the problem that the unambiguous low-frequency spectrum has very weak energy, which may cause a decrease in localization performance due to insufficient peak energy at the sound source location. Whether to use some spectrum enhancement techniques or other unambiguous methods to extend the scope of application is still a question worthy of further study. Theoretically, the angular spectrum-based method based on pairwise microphones has no additional restrictions on the number of microphones or the topology. In this paper, the applied microphone array consists of only two microphones to focus on the improvement of the MS spectrum. More microphone pairs of different topologies (e.g. planar, spherical) can be added for some specific applications in the following research.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by The National Natural Science Foundation of China (Nos. 61171167 and 61401203) and The Scientific Research Foundation of Jinling Institute of Technology (No. JIT-040520400101).

References

[1] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 909–912, Istanbul, Turkey, June 2000.

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

[3] A. Brutti, M. Ravanelli, P. Svaizer, and M. Omologo, "A Speech Event Detection and Localization Task for Multiroom environments," in *Proceedings of the 2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pp. 157–161, Villers-les-Nancy, France, May 2014.

[4] K. Wu and A. W. H. Khong, *Sound Source Localization and Tracking*, Springer International Publishing, New York, NY, USA, 2016.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[6] V. G. Reju, S. N. Soo Nqee Koh, and I. Y. Ing Yann Soon, "Underdetermined convolutive blind source separation via time-frequency masking," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 1, pp. 101–116, 2010.

[7] W. Wenyi Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 8, pp. 1913–1928, 2010.

[8] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 121–133, 2010.

[9] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

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

[11] A. Karbasi and A. Sugiyama, "A New DOA Estimation Method Using a Circular Microphone Array," in *Proceedings of the 15th European Signal Processing Conference*, pp. 778–782, Poznan, Poland, 2007.

[12] M. Jia, J. Sun, and C. Bao, "Real-time multiple sound source localization and counting using a sound field microphone," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 6, pp. 829–844, 2017.

[13] J. Dmochowski, J. Benesty, and S. Affes, "On spatial aliasing in microphone arrays," *IEEE Transactions on Signal Processing*, vol. 57, no. 4, pp. 1383–1395, 2009.

[14] V. V. Reddy, A. W. H. Khong, and B. P. Ng, "Unambiguous speech doa estimation under spatial aliasing conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2133–2145, 2014.

[15] 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 Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1079–1093, 2016.

[16] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[17] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.

[18] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 2, pp. 148–152, 1996.

[19] F. Nesta, P. Svaizer, and M. Omologo, *Independent Component Analysis and Signal Separation*, Springer Berlin Heidelberg, Berlin, Germany, pp. 290–297, 2009.

[20] F. Nesta and M. Omologo, "Generalized state coherence transform for multidimensional TDOA estimation of multiple sources," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, no. 1, pp. 246–260, 2012.

[21] F. Nesta and A. Brutti, "Tracking of multidimensional TDOA for multiple sources with distributed microphone pairs," *Computer Speech & Language*, vol. 27, no. 3, pp. 660–682, 2013.

[22] M. Cobos, J. J. Lopez, and D. Martinez, "Two-microphone multi-speaker localization based on a laplacian mixture model," *Digital Signal Processing*, vol. 21, no. 1, pp. 66–76, 2011.

[23] B. Loesch and B. Yang, "Online blind source separation based on time-frequency sparseness," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2009)*, pp. 117–120, Washington, DC, USA, April 2009.

[24] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *Proceedings of the 11th International Workshop on Acoustic Echo and Noise Control(IWAENC 2008)*, Seattle, WA, USA, September 2008.

[25] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[26] J. Li, X. Zhang, J. Tang, J. Cai, and X. Liu, "Audio magnetotelluric signal-noise identification and separation based on multifractal spectrum and matching pursuit," *Fractals*, vol. 27, no. 1, Article ID 1940007, 2019.

[27] Y. Fang and Z. Xu, "Multiple sound source localization and counting using one pair of microphones in noisy and reverberant environments," *Mathematical Problems in Engineering*, vol. 2020, no. 5, 12 pages, Article ID 8937829, 2020.

[28] B. Yang, H. Liu, C. Pang, and X. Li, "Multiple sound source counting and localization based on TF-wise spatial spectrum clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1241–1255, 2019.

[29] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 791–803, 2003.

[30] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 6, pp. 1429–1439, 2010.

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, PA, USA, 1991.

[32] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[33] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.