Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8937829, 12 pages
https://doi.org/10.1155/2020/8937829

Research Article

Multiple Sound Source Localization and Counting Using One Pair of Microphones in Noisy and Reverberant Environments

Yuzhuo Fang[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 2 April 2020; Revised 23 June 2020; Accepted 5 August 2020; Published 7 September 2020

Academic Editor: Emilio Insfran Pelozo

Copyright © 2020 Yuzhuo Fang and Zhiyong Xu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A multiple sound source localization and counting method based on an angular spectrum is proposed in this paper. Local signal-to-noise ratio tracking, onset detection, and a coherence test are introduced to filter the generalized cross-correlation angular spectrum in the time-frequency domain for multiple sound source localization and counting in noisy and reverberant environments. Then, dual-width matching pursuit is introduced to replace peak search as the method of localization and counting. A comprehensive comparison of two statistical indicators, mean precision and mean absolute estimated error, indicates that the proposed localization and counting algorithm using both the filtered angular spectrum and dual-width matching pursuit method is more robust and accurate than the classic counterpart, especially in environments with low signal-to-noise ratio, strong reverberation, and abundant sound sources.

1. Introduction

In the field of array signal processsing, multiple sound source localization and counting is a critical issue for applications such as indoor conferences, building security, and virtual reality [1]. Multiple sound source localization is usually subject to ambient noise, reverberation caused by confined space and obstacles, and mutual interference between multiple sound signals [2]. Many heuristic algorithms have been studied over the past two decades to suppress the effects of these negative factors and achieve robust and accurate performance. The main branch of research is based on time-frequency (TF) processing, which can be divided into three categories: clustering [3–6], histogram [7–10], and angular spectrum [6, 11–14].

The first category clusters the TF bins associated with each sound source on the basis of a criterion, such as interphase difference (IPD) [3] and direction estimation of mixing matrix (DEMIX) [4]. This method which directly achieves localization and counting results by iterative processing is sensitive to the initial clustering parameter [6]. In the second category, the weighted histogram using one pair of microphones is computed in [7] to solve the underdetermined problem. By using a relatively more complex topology, such as a unit circular array [9] and a soundfield microphone [10], the redundancy information of the circular integrated cross spectrum (CICS) [9] and the smoothed histogram [10] can improve the localization and counting accuracy, while expanding the estimated range of direction of arrival (DOA) from $[0, 180]^\circ$ to $[0, 360]^\circ$. These histogram-based algorithms are sensitive to the spatial aliasing ambiguity in some widely spaced microphone arrays because of the local TF computation [5]. In this paper, we focus on the third category, angular spectrum, which consists of two steps: (i) angular spectrum construction and (ii) sound source localization and counting.

First, an enumerated function related to all possible DOAs in each TF bin is accumulated to construct the angular spectrum [6], such as classical generalized cross-correlation phase transform (GCC-PHAT, simplified as GCC in the following) [11, 12] and the kernel density estimator (KDE) [13, 14]. GCC which achieves a certain degree of

antinoise performance through the phase weighting of the cross-correlation function is robust to the spatial aliasing problem [6, 15]. However, in environments with the coexistence of noise, reverberation, and mutual interference between multiple sound sources, the performance deteriorates substantially because of the limitations of the ideal single-source propagation model [16]. KDE has better antireverberation performance than GCC when the number of sound sources is relatively small (e.g., 2) [17]. The spatial aliasing ambiguity can be suppressed by the embedded frequency-dependent weighting factors in the kernel function, but the suppression is sensitive to the choice of kernel bandwidth [18].

Second, the DOAs and the corresponding number of multiple sound sources are obtained through the exhaustive search of the optimal parameter value from the angular spectrum. Traditional peak search (PS) [19], which is based on single-point peak amplitude, implements source localization and counting by comparing the peak amplitude with the cut-off threshold. The cut-off threshold becomes adaptive by using the previous peak [9]. Because the angular spectrum is seriously distorted in adverse environments, the estimated DOA when using PS may have a large offset, resulting in instability. Matching pursuit (MP) [20] can be used to improve the performance of PS by calculating the maximum inner product, but the choice of matching structure and atom width needs careful consideration. On the basis of source contributions, iterative contribution removal (ICR) [15] filters out the TF bins associated with the current estimated sound source during each iteration and then reconstructs the angular spectrum from the remainder to the next source localization. The reconstruction can restrain the distortion of the angular spectrum, but the search of the TF bins is computationally expensive, and incorrect previous source localization may exert considerable influence on the following iterative process.

In this paper, we improve each step of the angular spectrum-based algorithm. GCC is used because of its applicability to any microphone spacing [6, 15]. The main innovations are as follows. In the angular spectrum construction step, three TF filtering modules, local signal-to-noise ratio (SNR) tracking [15, 21], onset detection [22, 23], and coherence test [24, 25], are introduced to extract the TF bins that are less disturbed by noise, reverberation, and mutual interference between sound sources, respectively. The filtered GCC is termed GCCTF. In the localization and counting step, PS with the single-point amplitude is replaced by MP [8, 20] with the inner product of the atom from the perspective of contribution removal. The dual-width structure [9] and the source merging module are used to improve the iteration efficiency and to remove the repeatedly estimated results, respectively.

The remainder of this paper is organized as follows. In Section 2, the classic GCC is constructed based on the signal propagation model of multiple sound sources; then, three TF filtering modules are introduced to produce the GCCTF spectrum. In Section 3, the dual-width structure and the source merging module are presented during the description of MP. Numerical comparisons between the proposed GCCTF-MP algorithm and the classical ones are given in Section 4. Finally, Section 5 concludes the paper.

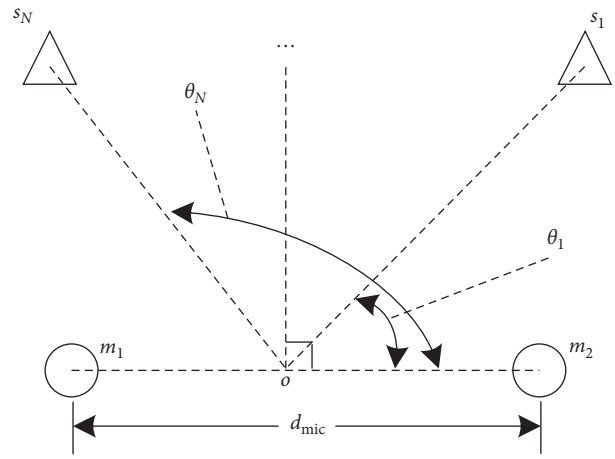

Figure 1: Propagation model of multiple sound sources.

2. GCC Angular Spectrum and TF Filtering

2.1. Signal Propagation Model of Multiple Sound Sources. The propagation model of multiple sound sources is shown in Figure 1. In the sound field space composed of several independent sound sources $s_1, \ldots, s_N$, there is a pair of omnidirectional microphones m_1 and m_2 with spacing d_{mic}. The DOA of the sound source, $\theta_1, \ldots, \theta_N \in [0, 180]°$, is defined in an anticlockwise manner, with 90° being the direction perpendicular to the line connecting the pair of microphones.

Ω_θ is indicated as the set of DOAs. The elements of Ω_θ and the corresponding number $N = \text{card}(\Omega_\theta)$ are unknown, where $\text{card}(\cdot)$ is the operator used to measure the number of elements in the set. In the approximately far field, the signal propagation model of multiple sound sources can be expressed as

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

where $x_m(i), m = 1, 2$, denotes the observed signal of the m-th microphone, $\mathbf{h}_{m,n} = [h_{m,n}(0), \ldots, h_{m,n}(L_h - 1)]^T$ denotes the impulse response between the n-th sound source and the m-th microphone, $\mathbf{s}_n(i) = [s_n(i), \ldots, s_n\ (i - L_h + 1)]^T$ denotes the discrete time signal vector of the n-th sound source with sampling rate f_s, and $w_m(i)$ is additive white Gaussian noise independent of the sources and the impulse responses.

When L_{FFT} points short-time Fourier transform (STFT), the expression in the discrete TF domain can be obtained as

$$X_m(r,k) = \sum_{n=1}^{N} H_{m,n}^T(r,k) S_n(r,k) + W_m(r,k), \tag{2}$$

where $X_m(r,k)$ and $S_n(r,k)$ denote the STFT coefficients of observed signal $x_m(i)$ and sound source signal $s_n(i)$ corresponding to the r-th frame and the k-th discrete frequency,

respectively; $W_m(r,k)$ denotes the additive complex noise; and $H_{m,n}(k)$ is the transfer function between the n-th sound source and the m-th microphone. Under the assumption of diffuse reverberation [26, 27], $H_{m,n}(k)$ can be decomposed as the direct wave component $H_{m,n}^{(\mathrm{D})}(k)$ and the reverberation component $H_{m,n}^{(\mathrm{R})}(k)$, that is,

$$\begin{aligned} H_{m,n}(k) &= H_{m,n}^{(\mathrm{D})}(k) + H_{m,n}^{(\mathrm{R})}(k), \\ &= \alpha_{m,n} e^{-j2\pi f_k T_{m,n}} + H_{m,n}^{(\mathrm{R})}(k), \end{aligned} \tag{3}$$

where $\alpha_{m,n}$ and $T_{m,n}$ denote the propagation attenuation and arrival time of the direct wave from the n-th sound source to the m-th microphone, respectively, and $f_k = (k f_s / L_{\mathrm{FFT}})$ denotes the frequency in the k-th frequency bin.

By inserting equation (3), equation (2) is expanded as

$$X_m(r,k) = \underbrace{\sum_{n=1}^{N} \alpha_{m,n} S_n(r,k) e^{-j2\pi f_k T_{m,n}}}_{X_m^{(\mathrm{D})}(r,k)} + \underbrace{\sum_{n=1}^{N} S_n(r,k) H_{m,n}^{(\mathrm{R})}(k)}_{X_m^{(\mathrm{R})}(r,k)} + W_m(r,k), \tag{4}$$

where $X_m^{(\mathrm{D})}(r,k)$ and $X_m^{(\mathrm{R})}(r,k)$ denote the direct wave component and the reverberation component of the observed signal $X_m(r,k)$, respectively.

3. GCC Angular Spectrum Construction

In an ideal environment, noise and reverberation do not exist, and the assumption of W-disjoint orthogonality (W-DO) is satisfied [28]. Both $X_m^{(\mathrm{R})}(r,k)$ and $W_m(r,k)$ in equation (4) are 0, and at most one sound source dominates the power in each TF bin. In this case, equation (4) can be simplified as

$$X_m(r,k) = \left(\alpha_{m,\eta(r,k)} S_{\eta(r,k)}(r,k)\right) e^{-j2\pi f_k T_{m,\eta(r,k)}}, \tag{5}$$

where $\eta(r,k) \in 1,\ldots,N$ denotes the index of the dominant sound source in each TF bin. Then, the IPD [3] which indicates the phase difference between the observed signals of the pair of microphones can be obtained as

$$\lambda(r,k) = \angle \frac{X_2(r,k)}{X_1(r,k)} = 2\pi f_k \tau_{\eta(r,k)} + \xi(r,k), \tag{6}$$

where $\angle$ is the operator to find the phase of a complex number, $\tau_{\eta(r,k)} = T_{2,\eta(r,k)} - T_{1,\eta(r,k)}$ is the time difference of arrival (TDOA) between the pair of microphones in each TF bin, and

$$\xi(r,k) = \left|2\pi f_k \tau_{\eta(r,k)}\right|_{2\pi}, \tag{7}$$

denotes the wrapping factor, where $|\cdot|_{2\pi}$ is the operator for obtaining the retained integer after $\mathrm{mod}(2\pi)$. The simulation results showed that wider microphone spacing can bring better resolution for localization [17]. However, the widening of d_{mic} is bound to break the limit of half the minimum wavelength $\lambda_{\min}$, thus making $\xi(r,k) \neq 0$, resulting in spatial aliasing ambiguity.

On the basis of the IPD obtained in equation (6), the local GCC related to the unknown DOA θ in each TF bin can be expressed as

$$\begin{aligned} \phi_{\mathrm{GCC}}(r,k,\theta) &= \Re\left(e^{\lambda(r,k)} \cdot e^{-j2\pi f_k \tau(\theta)}\right), \\ &= \Re\left(e^{2\pi f_k \left(\tau_{\eta(r,k)} - \tau(\theta)\right)} \cdot \underbrace{e^{j2\pi\xi(r,k)}}_{1}\right), \end{aligned} \tag{8}$$

where $\tau(\theta) = -d_{\mathrm{mic}} \cos((\theta)/c)$ (c denotes the atmospheric sound velocity) and $\Re(\cdot)$ denotes the real part of the complex argument. Because $e^{j2\pi\xi(r,k)} \equiv 1$, as shown in equation (8), the wrapping factor can be eliminated. By accumulating the local GCC in equation (8) across all TF bins, the GCC angular spectrum [6] can be obtained as

$$\Phi_{\mathrm{GCC}}(\theta) = \sum_{(r,k)} \phi_{\mathrm{GCC}}(r,k,\theta), \quad \theta \in \Omega_\theta^{(\mathrm{total})}, \tag{9}$$

where $\Omega_\theta^{(\mathrm{total})}$ denotes the linear space of $[0,180]^\circ$ with angle grid $\theta_{\min}$ and $\Omega_{\mathrm{TF}}^{(\mathrm{total})}$ denotes the set of all TF bins.

4. TF Filtered GCC Angular Spectrum

In practice, noise and reverberation are inevitable, and sources are more likely to overlap in the TF domain when the number of simultaneously occurring sound sources increases [9, 10]. Therefore, three modules, local SNR tracking, onset detection, and coherence test, are used to extract the TF bins that are less affected by the above problem in the GCC angular spectrum construction step. A block diagram of GCC and its filtered variant, GCCTF, is shown in Figure 2.

Local SNR tracking: the stronger noise contained in the observed signal will generate a higher angular spectrum floor with a lower recognition degree. So it is necessary to consider the enhancement of the observed signal which is often realized by tracking the noise floor [21]. In this paper, a simple SNR tracking method is used to obtain the SNR in each TF bin, namely, the local SNR [15], which can be expressed as

$$\gamma_{\mathrm{ST}}(r,k) = \min\left\{ \log_{10}\left(\frac{\left|X_m(r,k)\right|^2}{P_{\mathrm{W}}(r,k)} - 1 \right) \middle| \, m = 1,2 \right\}, \tag{10}$$

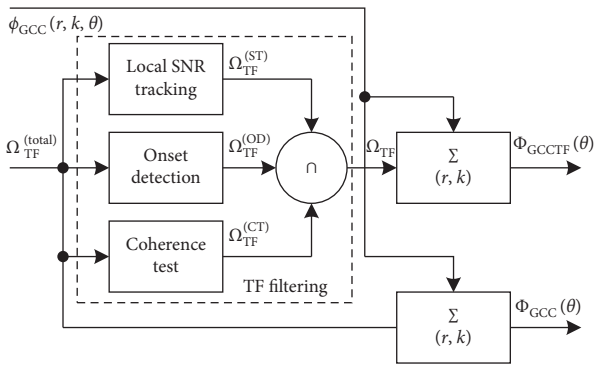

Figure 2: A block diagram of GCC and GCCTF.

where $P_W(r,k)$ denotes the local noise power and $\min\{\cdot\}$ denotes the operator used to obtain the minimum of the set.

Assume an ideal case that the whole observed signal starts with a section of pure noise. Then, the first L_W frames are used to measure the initial local noise power. L_W should not be set too long since the local noise power changes slowly over time. It is usually set to 2 or 3 empirically. The initial local noise power can be expressed as

$$P_W(r,k) = \left(\frac{1}{L_W}\right)\sum_{1}^{L_W}\left|X_m(r,k)\right|^2, \quad r = 1,\ldots,L_W. \tag{11}$$

Then, the local noise power increases slowly during signal frames and decreases slowly during noise frames. It can be expressed as

$$P_W(r+L_W,k) = \begin{cases} (1-\beta_{ST})P_W(r+L_W-1,k), & \text{noise frame,} \\ (1+\beta_{ST})P_W(r+L_W-1,k), & \text{signal frame,} \end{cases} \tag{12}$$

where β_{ST} is the updating factor.

The TF bins with local SNR above a user-defined threshold Γ_{ST} are extracted. The set of TF bins that satisfy the local SNR tracking module can be expressed as

$$\Omega_{TF}^{(ST)} = \{(r,k) \mid \gamma_{ST}(r,k) > \Gamma_{ST}\}. \tag{13}$$

Onset detection: in real environments, strong reverberation will bring serious angular spectrum distortion, which causes incorrect localization and counting results. Regardless of the special case that a new sound is very weak in each frequency band, the onset of a new sound, which is related to the direct wave component $X_m^{(D)}(r,k)$, is often accompanied by a sudden rise in signal amplitude (energy) within some frequency bands [22]. To detect this rise, the parameter of onset detection in each TF bin is set as follows [23]:

$$\gamma_{OD,m}(r,k) = g\left(\left|X_m(r,k)\right| - \left|X_m(r-1,k)\right|\right), \tag{14}$$

where $g(\cdot) = ((\cdot + |\cdot|)/2)$ is the half-wave rectifier function.

For the m-th microphone, once the peak of $\gamma_{OD,m}(r,k)$ is detected in each frequency band, the $X_m(r,k)$ corresponding to the same TF bin is considered as the onset. The threshold is set as $|X_m(r,k)|$ and is then gradually attenuated as the time frame r moves forward:

$$\Gamma_{OD,m}(r,k) = \begin{cases} \left|X_m(r,k)\right|, & \text{onset,} \\ \beta_{OD}\cdot\Gamma_{OD,m}(r-1,k), & \text{otherwise,} \end{cases} \tag{15}$$

where β_{OD} is the decaying factor empirically decided according to the experimental environment $\beta_{OD} < 1$. The set of TF bins that satisfy the onset detection module can be expressed as

$$\begin{aligned} \Omega_{TF}^{(OD)} = &\left\{(r,k) \mid X_1(r,k) \geq \Gamma_{OD,1}(r,k)\right\}, \\ &\cap \left\{(r,k) \mid X_2(r,k) \geq \Gamma_{OD,2}(r,k)\right\}. \end{aligned} \tag{16}$$

Coherence test: in practice, when the number of simultaneously occurring sound sources increases, the probability of sources with comparable power overlapping in the same TF bin increases. The W-DO assumption cannot be strictly met. So the assumption is appropriately relaxed: when accumulating the angular spectrum, there are TF bins with only one dominant sound source. Then, the coherence test module can be used to extract these TF bins effectively, which mitigates the effect of simultaneously occurring sources. The coherence test parameter is set as follows [25]:

$$\gamma_{CT}(r,k) = \left|\frac{E\left(X_1(r,k)X_2^*(r,k)\right)}{\sqrt{E\left(X_1(r,k)X_1^*(r,k)\right)}\sqrt{E\left(X_2(r,k)X_2^*(r,k)\right)}}\right|, \tag{17}$$

where $(\cdot)^*$ denotes the complex conjugate operator, $E(\cdot)$ denotes the average expectation of the $2C+1$ consecutive time frames, and

$$\begin{aligned} E\left(X_m(r,k)X_{m'}^*(r,k)\right) = &\frac{1}{2C+1}\sum_{r_1=r-C}^{r+C} X_m(r',k) \\ &\cdot X_{m'}^*(r',k), \quad m,m' \in \{1,2\}. \end{aligned} \tag{18}$$

TF bins with $\gamma_{CT}(r,k)$ above the user-defined threshold Γ_{CT} are considered to contain only one dominant sound source. Then, the corresponding set can be expressed as

$$\Omega_{TF}^{(CT)} = \left\{(r,k) \mid \gamma_{CT}(r,k) > \Gamma_{CT}\right\}. \tag{19}$$

On the basis of the three TF filtering modules (local SNR tracking, onset detection, and coherence test), combined with the derivation of GCC, the GCCTF angular spectrum can be obtained as

$$\Phi_{GCCTF}(\theta) = \sum_{(r,k)\in\Omega_{TF}} \phi_{GCC}(r,k,\theta), \tag{20}$$

where $\Omega_{TF} = \Omega_{ST} \cap \Omega_{OD} \cap \Omega_{CT}$ is the set of TF bins after TF filtering.

5. Dual-Width Matching Pursuit Method

After the angular spectrum construction in the previous section, the angular spectrum vector with the length $\text{card}(\Omega_\theta^{(total)})$ is used to realize the multiple sound source localization and counting. It can be expressed as

$$\Phi_{\text{str}} = \left[\Phi_{\text{str}}(0), \ldots, \Phi_{\text{str}}(z\theta_{\min}), \ldots, \Phi_{\text{str}}(180)^{\circ}\right], \quad z \in \mathbb{N}, \tag{21}$$

where the subscript "str" represents the string "GCC" or "GCCTF." Without loss of generality, Φ_{str} is simplified to Φ for convenience.

PS realizes localization and counting through the extraction of the spectrum peak above the cut-off threshold Γ_{PS}, while MP uses the inner-product comparison and the iterative source contribution removal. Consider an atom with one signal pulse, which can be approximately seen as the basic unit of the angular spectrum vector. The set of all the atoms can be defined as

$$\Omega_{\mathbf{u}} = \left\{ \mathbf{u}^{\longrightarrow(q)} \,\middle|\, 0 \le q \le \text{card}\left(\Omega_{\theta}^{(\text{total})}\right) - 1, \quad q \in \mathbb{Z} \right\}, \tag{22}$$

where $\longrightarrow$ denotes the operator of a circular shift to the right, $\mathbf{u}^{\longrightarrow(q)}$ denotes the row vector of $\mathbf{u}$ shifted to the right by q bits, and $\mathbf{u}$ can be expressed as

$$\mathbf{u} = \mathbf{u}^{\longrightarrow(0)} = \frac{\mathbf{v}^{\longrightarrow(-Q)}}{\left\|\mathbf{v}^{\longrightarrow(-Q)}\right\|}, \tag{23}$$

where $\|\cdot\|$ denotes the ℓ_2-norm operator of a vector, $\mathbf{v}^{\longrightarrow(-Q)}$ denotes the row vector of $\mathbf{v}$ shifted to the left by Q bits, and $\mathbf{v}$ can be expressed as

$$\mathbf{v} = \mathbf{v}^{\longrightarrow(0)} = \left[\mathbf{b}, \underbrace{0, \ldots, 0, \ldots, 0}_{\text{card}\left(\Omega_{\theta}^{(\text{total})}\right) - (2Q+1)}\right]^{T}, \tag{24}$$

where $\mathbf{b}$ is a Blackman window with width $2Q + 1$ and Q denotes one half of the window width $Q \in \mathbb{N}_{+}$.

The choice of Q must consider a compromise: in the same noisy and reverberant environment, an excessively wide width may incorrectly estimate the DOAs of sound sources with small angular intervals, and an excessively narrow width may reduce the iteration efficiency.

Therefore, a dual-width structure is proposed, where a narrower width is used to localize and a wider width is used to process the iterative contribution removal. To differentiate, subscripts "1" and "2" are added to parameters Q, $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{b}$ to indicate the atoms with narrower and wider window widths, respectively. A block diagram of the proposed MP method is shown in Figure 3, where the initial value $\boldsymbol{\Phi}(1)$ is set as Φ. The corresponding steps are as follows:

(1) DOA estimation: the inner-product of each atom with a narrower window width $\mathbf{u}_1^{\longrightarrow(q)}$ in set $\Omega_{\mathbf{u}_1}$ and the i-th angular spectrum vector $\boldsymbol{\Phi}(i)$ can be expressed as

$$p(q, i) = \left\langle \mathbf{u}_1^{\longrightarrow(q)}, \boldsymbol{\Phi}(i) \right\rangle, \tag{25}$$

where $\langle\cdot\rangle$ denotes the inner-product operator. Then, through the search of the maximum inner product, the estimated DOA in the i-th iteration can be obtained as

$$\begin{aligned} \hat{\theta}(i) &= \arg\max_{q} p(q, i) \cdot \theta_{\min}, \\ &= \dot{q}(i) \cdot \theta_{\min}, \end{aligned} \tag{26}$$

where $\dot{q}(i)$ is the shifted bit of $\mathbf{u}_1$ when the maximum inner product in the i-th iteration is obtained.

(2) Contribution measurement: on the basis of the maximum inner product in equation (25) and the corresponding atom with wider window width, the contribution vector of the estimated sound source in the i-th iteration can be measured as

$$\boldsymbol{\kappa}(i) = p(\dot{q}(i), i) \cdot \mathbf{u}_2^{\longrightarrow \dot{q}(i)}. \tag{27}$$

(3) Stop judgement: give two conditional expressions:

$$\frac{\kappa_{\text{sum}}(i)}{\kappa_{\text{sum}}(1)} < \Gamma_{\text{MP}}, \tag{28}$$

$$i > I_{\max}, \tag{29}$$

where Γ_{MP} is the user-defined threshold, $I_{\max}$ denotes the maximum number of iterations, and $\kappa_{\text{sum}}(i) = \text{sum}(\boldsymbol{\kappa}(i))$ denotes the contribution corresponding to the contribution vector $\boldsymbol{\kappa}(i)$ in equation (27), where $\text{sum}(\cdot)$ denotes the operator for the summation of the vector elements. If either equation (28) or equation (29) is satisfied, the loop stops.

(4) Residual calculation: after removing the contribution vector $\boldsymbol{\kappa}(i)$ from the angular spectrum vector $\Phi(i)$ in the i-th iteration, the residual used in the $(i + 1)$-th iteration can be calculated as

$$\boldsymbol{\Phi}(i + 1) = \boldsymbol{\Phi}(i) - \boldsymbol{\kappa}(i). \tag{30}$$

Set the number of iterations when the loop stops as I_{loop}; then, the set of the estimated DOAs after the loop body can be represented as

$$\Omega_{\widehat{\theta}_I} = \left\{ \widehat{\theta}(i) \,\middle|\, 1 \le i \le I_{\text{loop}} - 1, \quad i \in \mathbb{Z} \right\}. \tag{31}$$

Due to the limited window width, the contribution of a certain sound source may not be completely removed from the angular spectrum in each iteration, resulting in closely located sound sources. Extra counts generated by these sources will cause some deviations in counting results. Thus, the postprocessing step called source merging is used to merge the closely located sources into only one source in case that redundantly estimated sound sources are counted.

(5) Source merging: set $A_{\min}$ as the minimum angular interval. Any two estimated DOAs whose angular interval is smaller than $A_{\min}$ should be merged into

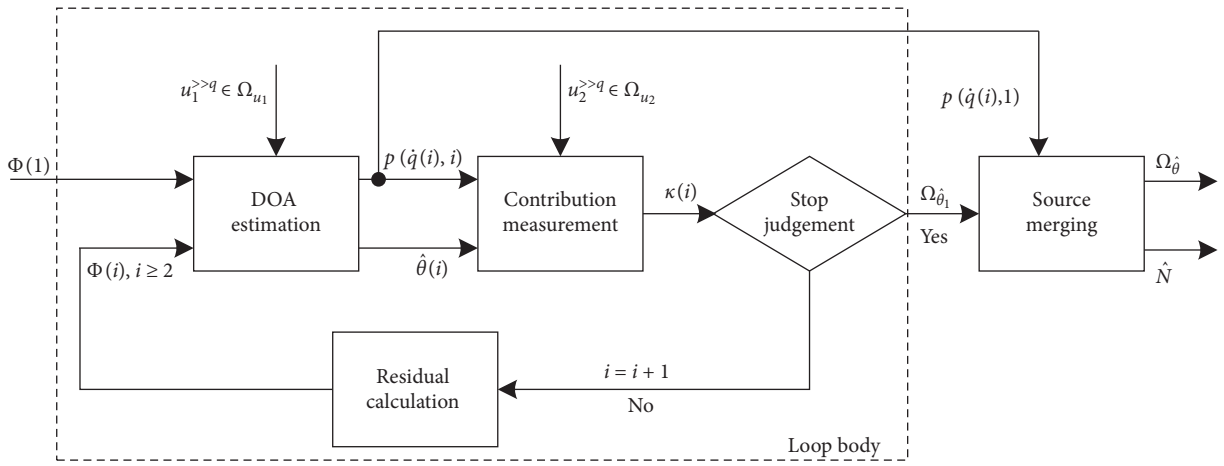

Figure 3: A block diagram of the MP method.

one sound source according to their corresponding initial inner products. If $|\widehat{\theta}(i') - \widehat{\theta}(i)| < A_{\min}$, where $\widehat{\theta}(i)$ and $\widehat{\theta}(i')$ are two estimated DOAs, the source merging process can be expressed as

$$\begin{cases} \hat{\theta}(i) = \hat{\theta}(i'), & p(\dot{q}(i), 1) > p(\dot{q}(i'), 1), \\ \hat{\theta}(i') = \hat{\theta}(i), & \text{otherwise.} \end{cases} \tag{32}$$

Then, the closely located sources are merged, and the final localization and counting results can be obtained as

$$\begin{cases} \Omega_{\widehat{\theta}} = \left\{\widehat{\theta}_n \mid n = 1, \ldots, \widehat{N}\right\}, \\ \widehat{N} = \text{card}\left(\Omega_{\widehat{\theta}}\right), \end{cases} \tag{33}$$

where $\Omega_{\widehat{\theta}}$ is the set of the final estimated DOAs, $\widehat{\theta}_n$ is the n-th element of $\Omega_{\widehat{\theta}}$, and $\widehat{N}$ is the number of the estimated DOAs.

6. Numerical Analysis

To verify the performance of the proposed GCCTF angular spectrum and dual-width MP method, the image-source model [27, 29] is used to generate the observed data, where the room size is 8.5 m × 7.5 m × 3 m and the sound velocity c is 344 m/s. The plane schematic diagram of the room is shown in Figure 4 where the heights of the microphones and sources are all set to 1.3 m. A pair of omnidirectional microphones m_1 and m_2 parallel to the x-axis is located at the center of room o with spacing $d_{\text{mic}} = 0.8$ m. N sound sources $s_1, \ldots, s_n$ are distributed on a semicircle with o as the centroid and microphone-source distance d_{ms} as the radius.

The DOA distribution when the true number of sources N varies from 2 to 6 is presented in Table 1. 8 male and 8 female voices taken from the TIMIT dataset are used as sound sources [30], with a sampling rate $f_s = 16$ kHz. The total number of simulations I_{sim} is set to 200. In each simulation, N segments of length 1.024 s from different voices are randomly extracted and then preprocessed to have the same average power.

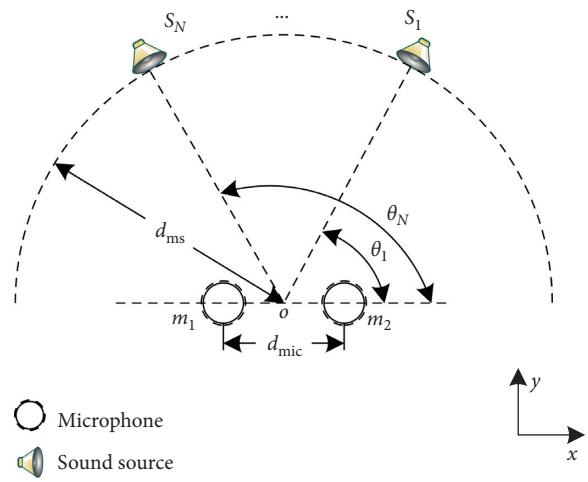

Figure 4: Plane schematic diagram of all the microphones and sources in the room.

Table 1: Number of sources N versus DOA.

N	DOA (°)
2	45, 135
3	15, 45, 135
4	15, 45, 135, 165
5	15, 45, 75, 135, 165
6	15, 45, 75, 105, 135, 165

From equation (1), the observed signal is produced through convolution of the source signal with the impulse response generated by the image-source model and then added with white Gaussian noise with

$$\text{SNR} = 10 \log_{10} \frac{(\overline{P}_1 + \overline{P}_2)}{2\overline{P}_{\text{W}}}, \tag{34}$$

where $\overline{P}_1$, $\overline{P}_2$, and $\overline{P}_{\text{W}}$ denote the average power of the two microphones m_1 and m_2 and the additive noise, respectively. We discuss the performance in three scenarios with different

reverberation times RT_{60} and d_{ms}: (i) $RT_{60} = 0.2$ s, $d_{ms} = 2$ m; (ii) $RT_{60} = 0.5$ s, $d_{ms} = 2$ m; and (iii) $RT_{60} = 0.5$ s, $d_{ms} = 3.5$ m. Direct to reverberation ratio (DRR) which indicates the reverberation level is defined as follows [31]:

$$\mathrm{DRR}_{m,n} = 10\log_{10}\frac{\sum_i \left(h_{m,n}^{(\mathrm{D})}(i)\right)^2}{\sum_i \left(h_{m,n}^{(\mathrm{R})}(i)\right)^2}, \tag{35}$$

where the numerator and denominator of the logarithmic function represent the total power of the direct wave component and the reverberation component of the impulse response between the n-th sound source and the m-th microphone, respectively. Then, the average DRRs of all the sources when $N = 6$ in scenario (i), scenario (ii), and scenario (iii) are 3.19 dB, −1.87 dB, and −6.60 dB, respectively.

7. Comparison of Angular Spectrum Recognition Degree

The parameter configuration in the angular spectrum construction step is presented in Table 2, where γ_{ST} is 5 dB, as suggested in [15], and other parameters are set empirically through the previous experiment. Figure 5 shows the normalized angular spectra of GCC and GCCTF when $N = 2$ and SNR = 10 dB from scenario (i) to scenario (iii) in a single simulation, where the arrows in each subfigure indicate the true DOAs. The peak amplitudes of the two angular spectra are almost the same as the true DOAs in scenario (i) and scenario (ii). The local SNR tracking module in GCCTF can efficiently reduce the angular spectrum floor so that the peak amplitude deviations from the true DOAs are lower than those of GCC. In scenario iii, the growth of d_{ms} makes the distance between the direct wave peak and the first reverberant peak shorter, which results in the aggravation over the reverberation level. The spectrum floor is higher than scenario (ii), and the false peak marked by "x" has a high amplitude compared to the true DOA in GCC, resulting in an incorrectly estimated DOA. The onset detection module in GCCTF can prevent this phenomenon so that GCCTF can retain the correct estimation.

GCCTF forms a more recognizable angular spectrum than GCC from scenario (i) to scenario (iii) in Figure 5. To quantitatively indicate the recognition degree from a statistical perspective, we introduce the following mean precision (MPRE), which can be expressed as follows [17, 18]:

$$\mathrm{MPRE} = \frac{1}{I_{\mathrm{sim}}}\sum_{i=1}^{I_{\mathrm{sim}}}\frac{N_{\mathrm{co}}(i)}{N_{0.2}(i)}, \tag{36}$$

where $N_{\mathrm{co}}(i)$ and $N_{0.2}(i)$ denote the number of correctly estimated DOAs and the spectrum peaks whose normalized amplitudes are greater than 0.2 in the i-th simulation, respectively. The criterion for judging whether the DOA is estimated correctly is that the interval between the true value and the estimated value is less than $A_{\min}$, which is 10°, as suggested in [9, 15].

Table 2: Parameter configuration in the angular spectrum construction step.

Parameter class	Name or notation	Value
STFT	Frame length	512
	Frame shift	256
	L_{FFT}	512
Angular spectrum	$\theta_{\min}$	0.5°
Local SNR tracking	L_{W}	3
	Γ_{ST}	5 dB
	β_{ST}	0.01
Onset detection	β_{OD}	0.45
Coherence test	C	2
	Γ_{CT}	0.8

Figure 6 shows the MPREs of GCC and GCCTF versus SNR from scenario (i) to scenario (iii). In each subfigure of Figure 6, MPREs improve with the growth of SNR, which is mainly due to the fact that the reduction of noise floor means the recognition degree improvement of angular spectrum. Horizontal comparison of all the subfigures in Figure 6 shows that MPREs tends to decrease from scenario (i) to scenario (iii), which is mainly due to the fact that the deterioration of reverberation will strengthen the effect of false peaks and thus decrease the recognition degree. GCCTF performs better than GCC under the same noisy and reverberant environment, which shows that TF filtering can indeed bring better recognition degree and then bring a positive impact on the subsequent localization and counting performance.

8. Comparison of the Localization and Counting Performances

The mean absolute estimated error (MAEE) is used to measure the localization and counting performance [9, 10]:

$$\mathrm{MAEE} = \frac{1}{I_{\mathrm{sim}}}\sum_{i=1}^{I_{\mathrm{sim}}}\frac{1}{N_{\max}(i)}\sum_{n=1}^{N_{\max}(i)}\left|\widehat{\theta}_n(i) - \theta_n(i)\right|, \tag{37}$$

where $\theta_n(i)$ and $\widehat{\theta}_n(i)$ denote the true and estimated DOA of the n-th sound source in the i-th simulation, respectively; and $N_{\max}(i) = \max(N(i), \widehat{N}(i))$, where $N(i)$ and $\widehat{N}(i)$ denote the numbers of true and estimated sound sources in the i-th simulation, respectively. Since $N_{(i)}$ is not greater than $N_{\max}(i)$, the excessively estimated DOAs are set to meet the following expression:

$$\left|\widehat{\theta}_n(i) - \theta_n(i)\right| = A_{\min}. \tag{38}$$

8.1. Choice of the Window Width in the MP Method. When using the MP method, the final localization and counting performance are affected by the choice of the dual-width. Through large numbers of simulations, we find that when Q_1, one half of the narrower window width, is chosen from 4 to 6, the maximum inner product can accurately determine the DOAs. In this case, the choice of Q_2, the wider counterpart, exhibits a regular change with N. To illustrate this

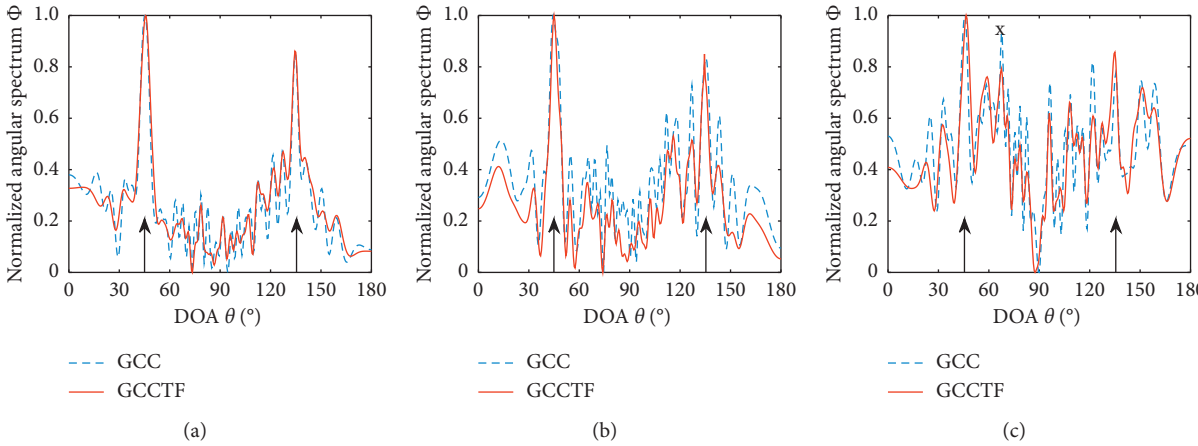

Figure 5: Angular spectra of GCC and GCCTF when $N = 2$ and SNR = 10 dB in three scenarios: (a) scenario (i); (b) scenario (ii); (c) scenario (iii).

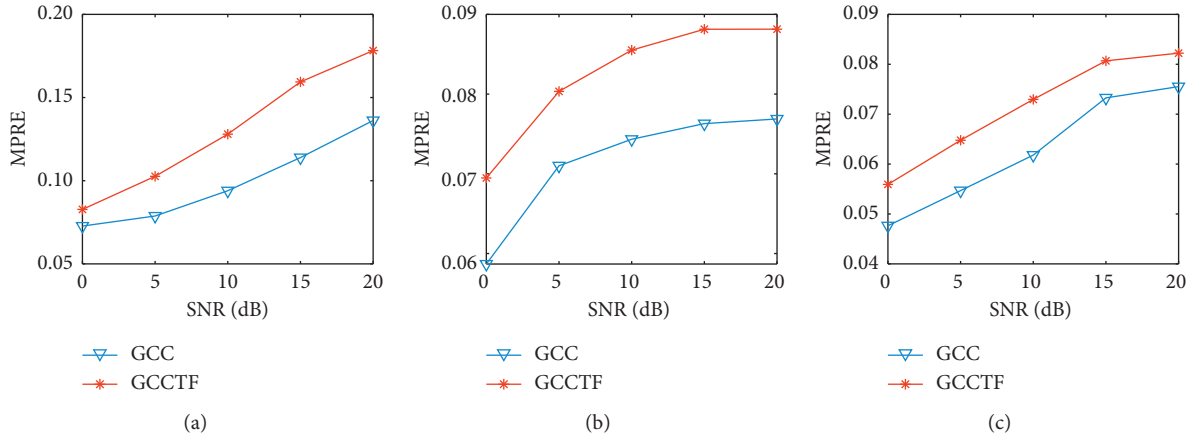

Figure 6: Angular spectra of GCC and GCCTF when $N = 2$ and SNR = 10 dB in three scenarios: (a) scenario (i); (b) scenario (ii); (c) scenario (iii).

regularity, without loss of generality, we present the MAEE with Q_1 fixed at 4 while Q_2 varies from 4 to 8 (in intervals of 2), where the GCCTF angular spectrum is used. Based on the comprehensive consideration of algorithm efficiency and accuracy, the maximum number of loops $I_{\max}$ is set to 10, and the cut-off thresholds Γ_{MP} in scenario (i), scenario (ii), and scenario (iii) are 0.51, 0.53, and 0.60, respectively.

Figures 7 and 8 show the MAEEs of GCCTF-MP in three scenarios when SNR = 15 dB and SNR = 5 dB, respectively. $Q_2 = 4$ (single-width) performs the best when $N = 3$ from scenario (i) to scenario (iii); then, the performance declines as the width increases.

In scenario (i), when N is 4, 5, and 6, the best performance is obtained with Q_2 set to 6, 8, and 10, respectively. Thus, in a low reverberant environment (e.g., scenario (i)), a wider window width improves the performance of MP when there are more sound sources.

The trends in scenario (ii) are similar, except that when SNR = 5 dB and $N = 4$, $Q_2 = 4$ performs the best. In scenario (iii), when SNR = 15 dB and $N = 5$, $Q_2 = 6$ performs the best; when SNR = 5 dB and N is 4, 5, and 6, the best performance is obtained with Q_2 set to 4, 4, and 6, respectively. Thus, increasing the window width results in sensitivity to the adverse environment with relatively strong reverberation and low SNR. However, the wider width Q_2 shows an improvement when there are more sound sources in the same noisy and reverberant environment.

8.2. Comparison of the Localization and Counting Performances. To provide a comprehensive comparison of the localization and counting performance, three algorithms, GCC-PS, GCCTF-PS, and GCC-MP, are obtained similarly to GCCTF-MP by combining the angular spectrum with the localization and counting method. Based on the previous analysis, where Q_1 is set to 4, Table 3 presents the parameter configuration when using the PS and MP methods from scenario (i) to scenario (iii).

Figures 9–11 show the MAEEs versus SNR when N varies from 3 to 6 in scenario (i), scenario (ii), and scenario (iii), respectively.

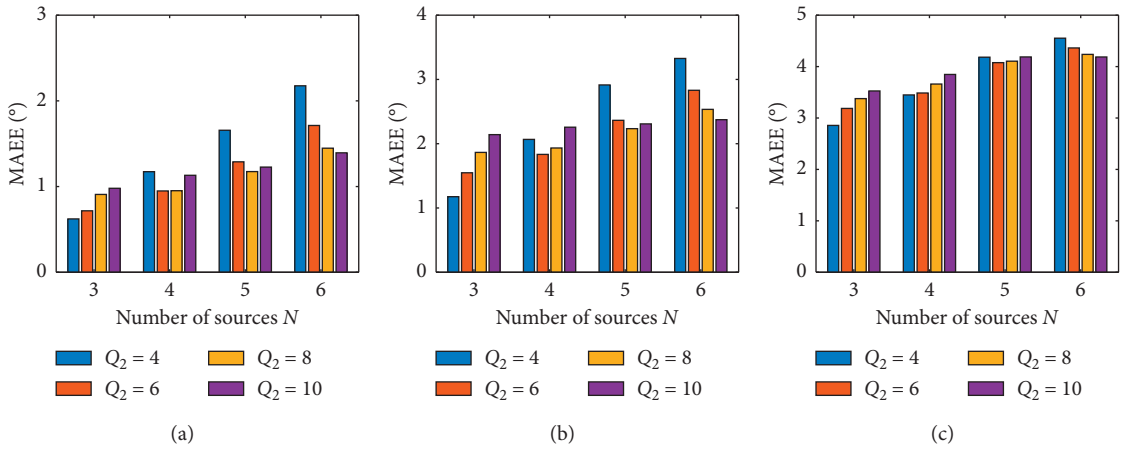

Figure 7: MAEEs of GCCTF-MP versus N when SNR = 15 dB in three scenarios: (a) scenario (i); (b) scenario (ii); (c) scenario (iii).

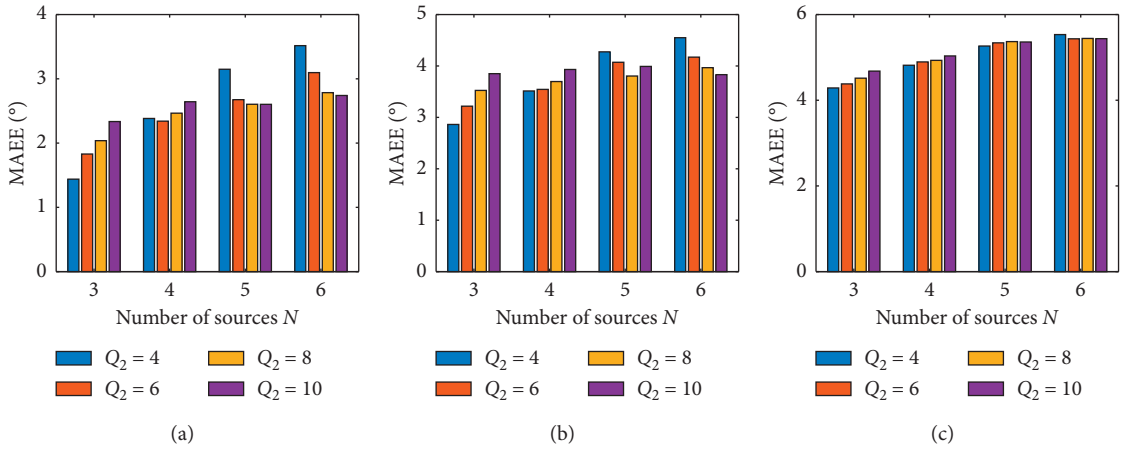

Figure 8: MAEEs of GCCTF-MP versus N when SNR = 5 dB in three scenarios: (a) scenario (i); (b) scenario (ii); (c) scenario (iii).

In Figure 9, the MAEEs ranges from 0.5° to 5.5°. As N increases, the MAEE increases due to aggravation of the mutual interference between sound sources. In each subfigure, the MAEE gradually decreases with increasing SNR. GCC-PS and GCCTF-MP perform the worst and best among the four algorithms. When SNR > 0 dB, the two algorithms using the MP method have lower MAEE than those using PS, and the two algorithms using the GCCTF angular spectrum have lower MAEE than those using GCC. When SNR = 0 dB, the MAEE of GCCTF-PS is lower than that of GCC-MP, which shows that the SNR tracking module can effectively reduce the TF bins with relatively low local SNR.

In Figure 10, medium reverberation degrades the performance of the four algorithms, with MAEE between 1° and 7°. The trend is similar to scenario (i), except when SNR ≤ 10 dB, GCCTF-PS performs better than GCC-MP. Compared to the turning point of 0 dB in scenario (i), in a medium reverberant environment (e.g., scenario (ii)), the onset detection module may have a positive impact on the final localization and counting.

In Figure 11, strong reverberation results in the worst performance for the four algorithms, with MAEE between 2° and 8°. The trend is similar to scenario (ii), except when $N > 4$ and SNR varies from 0 to 15 dB, GCCTF-PS performs better than GCC-MP. Compared to the turning point of 10 dB in scenario (ii), in a relatively strong reverberant environment (e.g., scenario (iii)), the coherence test module may have a positive impact on the final localization and counting.

Based on the results from Figures 9 to 11, GCCTF can provide a more robust angular spectrum than GCC

Table 3: Parameter configuration when using PS and MP.

Notation	Condition	Value
Q_2	Scenarios (i) and (ii); $N = 3$ Scenario (iii); $N = 3, 4$ Scenario (iii); $N = 5$; SNR ≤ 5 dB	4
	Scenarios (i) and (ii); $N = 4$ Scenario (iii); $N = 5$; SNR > 5 dB	6
	Scenarios (i) and (ii); $N = 5$	8
	Scenarios (i)–(iii), $N = 6$	10
Γ_{PS}	Scenario (i)	0.57
	Scenario (ii)	0.63
	Scenario (iii)	0.67
Γ_{MP}	Scenario (i)	0.51
	Scenario (ii)	0.53
	Scenario (iii)	0.60

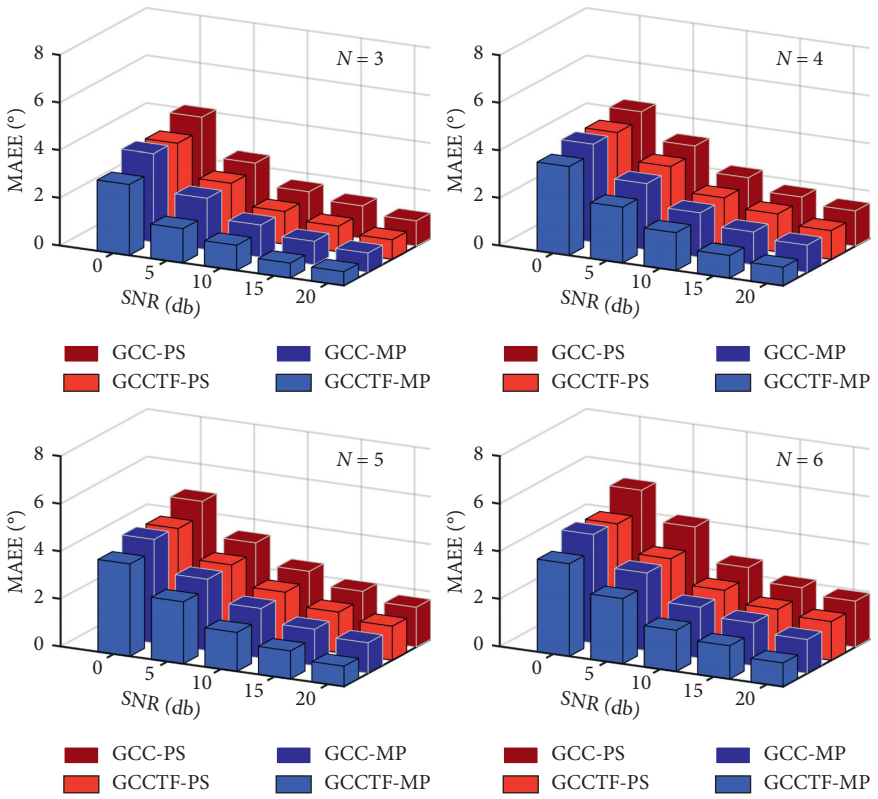

Figure 9: MAEEs of GCC-PS, GCCTF-PS, GCC-MP, and GCCTF-MP versus SNR when N varies from 3 to 6 in scenario (i).

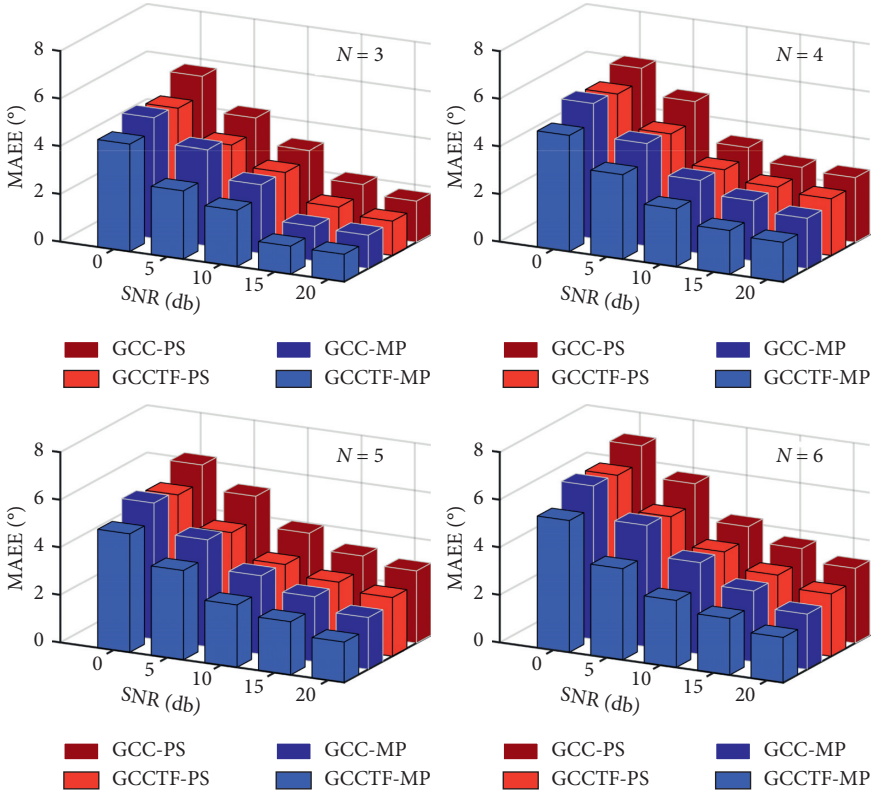

Figure 10: MAEEs of GCC-PS, GCCTF-PS, GCC-MP, and GCCTF-MP versus SNR when N varies from 3 to 6 in scenario (ii).

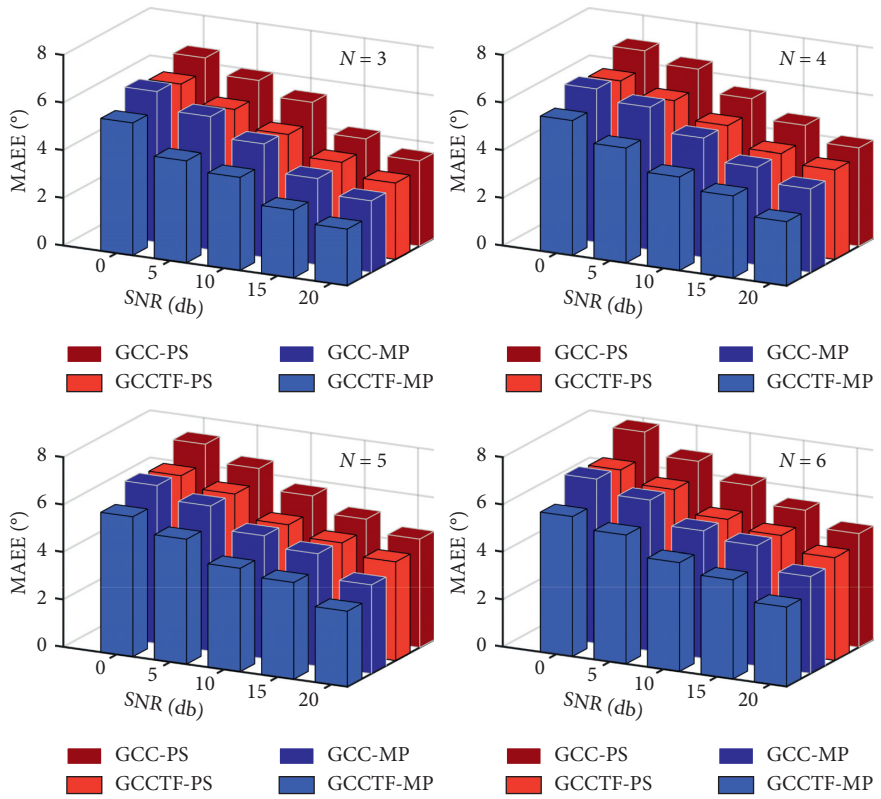

Figure 11: MAEEs of GCC-PS, GCCTF-PS, GCC-MP, and GCCTF-MP versus SNR when N varies from 3 to 6 in scenario (iii).

especially under the acoustically adverse environment. This is because that the TF bins seriously affected by low SNR, strong reverberation, and abundant sound sources have been removed by the three filtering modules. So the distortion of the angular spectrum can be effectively alleviated, and the increase in the false peak amplitude can be suppressed simultaneously. MP performs better than PS because PS may produce instable performance with severely distorted angular spectrum. GCCTF-MP, which uses both the GCCTF angular spectrum and MP method, is the most robust and accurate multiple sound source localization and counting algorithm among the four combined algorithms.

9. Conclusion

In this paper, GCCTF-MP, an algorithm for multiple sound source localization and counting, is proposed for noisy and reverberant environments. Three modules, local SNR tracking, onset detection, and a coherence test, are used to filter the GCC angular spectrum; the dual-width MP method is used to replace the amplitude comparison with the inner product and contribution removal. On the basis of the statistical indicators MPRE and MAEE, MP is shown to be a more accurate method than PS, and GCCTF is shown to be a more recognizable and robust angular spectrum, especially in environments with low SNR, strong reverberation, and abundant sound sources. The proposed GCCTF-MP, which uses both the GCCTF angular spectrum and MP method, is thus a robust and accurate multiple sound source localization and counting algorithm. Furthermore, we find that the final localization and counting performance when using the MP method is affected by the choice of the dual-width. A brief comparison when using a fixed narrower width and a different wider counterpart is presented. In practice, the environmental parameters are difficult to determine, and the number of sound sources is unknown. How to implement the width in an adaptive manner is a challenging problem that warrants further study.

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] K. Wu and A. Khong, "Sound source localization and tracking," *Context Aware Human-Robot and Human-Agent Interaction*, Springer International Publishing, Cham, Switzerland, 1st edition, 2016.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer International Publishing., Cham, Switzerland, 2008.

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

[4] 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.

[5] 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.

[6] 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.

[7] J. Escolano, N. Xiang, J. M. Perez-Lorenzo, M. Cobos, and J. J. Lopez, "A Bayesian direction-of-arrival model for an undetermined number of sources using a two-microphone array," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 742–753, 2014.

[8] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[9] 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.

[10] M. Jia, J. Sun, and C. Bao, "Real-time multiple sound source localization and counting using a soundfield microphone," *Journal of Ambient Intelligence & Humanized Computing*, vol. 8, no. 6, pp. 1–16, 2016.

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

[12] 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.

[13] 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.

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

[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] 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.

[17] Y. Fang and Z. Xu, "A robust algorithm for unambiguous TDOA estimation of multiple sound sources under indoor environment," *Journal of Electronics & Information Technology*, vol. 38, no. 5, pp. 1143–1150, 2016.

[18] 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.

[19] 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.

[20] 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.

[21] E. Yariv and M. David, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech & Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[22] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[23] S. Dixon, "Onset detection revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects*, Montreal, Quebec, Canada, September 2006.

[24] N. Tho, S. Zhao, and D. Jones, "Robust DOA estimation of multiple speech sources," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2014)*, Florence, Italy, May 2014.

[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," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2136–2147, 2008.

[26] T. Gustaffson, B. 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.

[27] 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.

[28] 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.

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

[30] J. Garofolo, L. Lamel, W. Fisher et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus Linguistic Data Consortium*, 1993.

[31] P. Naylor, N. Gaubitch, and E. Habets, "Signal-based performance evaluation of dereverberation algorithms," *Journal of Electrical & Computer Engineering*, vol. 2010, Article ID 127513, 5 pages, 2010.