An Iterative Approach to Source Counting and Localization Using Two Distant Microphones

Lin Wang, Tsz-Kin Hon, Joshua D. Reiss, and Andrea Cavallaro

***Abstract*—We propose a time difference of arrival (TDOA) estimation framework based on time-frequency inter-channel phase difference (IPD) to count and localize multiple acoustic sources in a reverberant environment using two distant microphones. The time-frequency (T-F) processing enables exploitation of the non-stationarity and sparsity of audio signals, increasing robustness to multiple sources and ambient noise. For inter-channel phase difference estimation, we use a cost function, which is equivalent to the generalized cross correlation with phase transform (GCC) algorithm and which is robust to spatial aliasing caused by large inter-microphone distances. To estimate the number of sources, we further propose an iterative contribution removal (ICR) algorithm to count and locate the sources using the peaks of the GCC function. In each iteration, we first use IPD to calculate the GCC function, whose highest peak is detected as the location of a sound source; then we detect the T-F bins that are associated with this source and remove them from the IPD set. The proposed ICR algorithm successfully solves the GCC peak ambiguities between multiple sources and multiple reverberant paths.**

***Index Terms*—GCC-PHAT, IPD, microphone array, source counting, TDOA estimation.**

I. Introduction

AD-HOC acoustic sensor networks composed of randomly distributed wireless microphones or hand-held smartphones have been attracting increasing interest due to their flexibility in sensor placement [1]–[5]. Sound source localization is a fundamental issue in ad-hoc acoustic sensor signal processing, with applications to tracking, signal separation and noise suppression [6]–[9], among others. An important problem in source localization is to estimate the number of active sources (source counting) [10], [11], because many multi-source localization [12], [13] and source separation algorithms [14], [15] require this information as input. Unlike the conventional regular structure of microphone arrays, the microphones in an ad-hoc arrangement can be far apart from each other, and therefore the inter-microphone delay can be high. Other challenges include multi-source and multi-path interaction [12], as well as spatial aliasing at high frequencies [10].

Dual-microphone techniques are crucial in an ad-hoc acoustic sensor network, since such a network can be seen as a combination of multiple microphone pairs and pairwise processing can increase the scalability, and thus also the robustness, of the network [16]. Counting and localizing multiple simultaneously active sources in real environments with only two long-distance microphones is usually an under-determined problem. Generally, source counting and localization can be achieved via time-frequency (T-F) clustering, which exploits the phase information of microphone signals, e.g., the linear variation of the inter-channel phase difference (IPD) with respect to frequency [12], [13]. The additive ambient noise at microphones will distort the desired phase information and degrade the source localization accuracy. The overlap of multiple sources contributing to the same T-F bin can also distort the desired phase information. T-F clustering approaches typically require that the arrangement of the microphones should satisfy the space sampling theorem, i.e., the inter-microphone distance should be smaller than half the wavelength (e.g., 4 cm for a sampling rate of 8 kHz), so that spatial aliasing will not occur [10]. This requirement is difficult to meet in an ad-hoc arrangement, and the long delay between two microphones may lead to wrapped IPD at high frequencies [12], [17]. This is the biggest challenge for T-F approaches. Another class of correlation-based approaches, e.g., generalized cross-correlation with phase transform (GCC-PHAT) [18], is robust to the phase wrapping problem. GCC-PHAT locates the source based on the peak of the generalized cross-correlation (GCC) function. However, the interaction between multiple sources and multiple reverberant paths generates a higher number of GCC peaks than the number of sources. This ambiguity raises a new challenge for estimating the number of sources.

In this paper, we propose a new framework for source counting and localization using two microphones, which can deal with scenarios with far-apart microphones (e.g., 0.15–6 m). The two main novelties of the paper are as follows. First, we merge the concept of the T-F IPD and the concept of GCC-PHAT. By using T-F weighting based on the SNR and the coherence measures, the nonstationarity and sparsity of audio signals can be exploited to improve the robustness to ambient noise and multiple sources. By using the GCC-PHAT cost function, the spatial ambiguity caused by a large inter-microphone distance can be solved. Second, we propose an iterative contribution removal (ICR) algorithm, which performs source localization and counting. The ICR algorithm successfully solves the peak ambiguities between multiple sources and multiple paths by exploiting the variation of IPD with frequency. In each iteration, the ICR algorithm detects one source from the GCC function and subsequently removes the T-F bins associated with this source for recalculating a new GCC function. In this way, source localization and source counting can be jointly achieved.

Manuscript received January 30, 2015; revised December 1, 2015 and January 29, 2016; accepted February 4, 2016. Date of publication April 1, 2016; date of current version April 29, 2016. This work was supported by the U.K. Engineering and Physical Sciences Research Council under Grant EP/K007491/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Woon-Seng Gan.

The authors are with the Centre for Intelligent Sensing, Queen Mary University of London, London E1 4NS, U.K. (e-mail: lin.wang@qmul.ac.uk; tsz.kin.hon@qmul.ac.uk; joshua.reiss@qmul.ac.uk; a.cavallaro@qmul.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2533859

TABLE I
SUMMARY OF SOUND SOURCE LOCALIZATION ALGORITHMS (M: NUMBER OF MICROPHONES; N: NUMBER OF SOURCES)

Approach		Reference	M	N	Source counting	Comments
Blind identification	Eigen decomposition	[7], [19]	≥ 2	≥ 1	No	High computational cost
	ICA	[20]–[22]				
Angular spectrum	GCC-PHAT	[18], [25]	2	1	No	Single source only
	SRP, SRP-PHAT	[25]–[28]	≥ 2	≥ 1		Extension of GCC-PHAT to more microphones
	MUSIC, ESPRIT	[29]–[32]				$M > N$
	Joint pitch-location estimation	[33]–[36]				Exploiting a harmonic model of source signals
T-F processing	Histogram	[37]–[39]	≥ 2	≥ 1	Yes	Closely spaced microphones only
	Clustering	[10]–[13] [40]–[47]				
	TF + Angular spectrum	Proposed	2			Suitable for far-apart microphones

The paper is organized as follows. Section II overviews related sound source localization and counting methods in the literature. Section III formulates the problem. The IPD-based source localization framework and the ICR source counting algorithm are proposed in Section IV and Section V, respectively. Performance evaluation is conducted in Section VI and conclusions are drawn in Section VII.

II. RELATED WORKS

Depending on how localization is achieved, source localization may also be referred to as time delay estimation, time difference of arrival (TDOA) estimation, and direction of arrival (DOA) estimation. We classify source localization algorithms into three groups, namely blind identification, angular spectrum, and T-F processing (Table I).

Blind identification algorithms estimate the acoustic transfer functions between the sources and the microphones, from which the DOAs of the sources can be easily obtained. Eigen-decomposition is a popular blind identification approach, which estimates the transfer function from the covariance matrix of microphone signals [7], [19]. Recently, independent component analysis based system identification has shown promising results [20]–[22]. One drawback of blind identification is its computational cost. For instance, the acoustic mixing filter can be several thousand taps long in reverberant scenarios and to estimate the large number of parameters of such a mixing system simultaneously and blindly can be computationally demanding.

Angular spectrum algorithms build a function of the source location which is likely to exhibit a high value at the true source location. Several approaches can be used to build such a function, e.g., GCC-PHAT, steered response power (SRP) and multiple signal classification (MUSIC). GCC-PHAT calculates the correlation function using the inverse Fourier transform of the cross-power spectral density function multiplied by a proper weighting function, and localizes the sound source from the peak of the GCC function [18], [25]. GCC-PHAT is suitable for far-apart microphones and has shown satisfactory results for a single source in reverberant but low-noise environments [23], [24]. A new challenge arises when applying GCC-PHAT to speech signals from two distant microphones on a short timescale (e.g., hundreds of milliseconds). The GCC function may have ambiguous peaks not only from the TDOA but also from the fundamental frequency (pitch) of the signal. SRP steers out beams and localizes high-energy sound sources. SRP-PHAT is an extension of the two-microphone GCC-PHAT to multiple pairs of microphones [25]–[28]. MUSIC is a subspace method for multi-DOA estimation, where the angular spectrum function is constructed from the steering vector of the candidate DOA and the eigenvector of the noise subspace [29]. Estimation of signal parameters by rotational invariance techniques (ESPRIT), another subspace-based algorithm, is more robust to array imperfections than MUSIC by exploiting the rotational invariance property in the signal subspace created by two subarrays, which are derived from the original array with a translation invariant structure [30]. Both MUSIC and ESPRIT were originally proposed for narrowband radar signals in anechoic scenarios and when the number of sensors is greater than the number of sources. The narrowband MUSIC and ESPRIT algorithms can also be extended to wideband applications [31], [32]. Pitch-location joint estimation approaches [33]–[36] assume a harmonic model of voiced speech and are robust against multi-source scenarios, since the location information helps improve pitch estimation for multiple sources while the pitch information helps distinguish sources coming from close locations.

T-F processing algorithms compute the DOA locally in each T-F bin and associate these DOAs to each source by means of a histogram or clustering [10]–[13], [37]–[45]. Several probability models have been proposed to model the distribution of multiple DOAs, such as Gaussian mixture model [12], Laplacian mixture model [12], [13] and Von Mises model [43]. The T-F approaches have been investigated intensively in recent years due to their source counting capability and application to underdetermined DOA estimation problems. Processing in the T-F domain allows one to exploit the nonstationarity and sparsity of audio signals to improve the robustness in noisy and multi-source scenarios. One drawback of the existing T-F approaches is that they are only suitable for closely spaced microphones since with widely spaced microphones the local DOA estimation becomes ambiguous due to spatial aliasing.

Most multi-source localization approaches need prior knowledge of the number of sources to operate properly. Among the three groups mentioned above, only the third considers how to estimate the number of sources. Source counting in T-F is

achieved by applying information criterion based model order selection [48], [49] when clustering the T-F bins [12], [39], [46], [47] or by counting the peaks of the DOA histogram [38]. However, T-F approaches typically require the inter-microphone distance to be smaller than half the wavelength, an assumption that is not satisfied in the applications we are interested in with far-apart microphones. An exception is [11] where spatial aliasing is avoided by applying clustering on the signal amplitude only, but this is not applicable to scenarios where the levels of the sources are similar. Thus, how to perform source counting and localization with large-distance microphones is still an open problem.

III. Problem Formulation

Consider $M = 2$ microphones, whose relative distance is d, and N physically static sound sources in a reverberant environment. N and the DOAs of sound sources $\theta_1, \ldots, \theta_N$ are all unknown. The sound direction is defined in an anti-clockwise manner with 90° being the direction perpendicular to the line connecting the two microphones. The microphone signals are synchronously sampled. The signal received at the mth microphone is

$$x_m(n) = \sum_{i=1}^{N} \boldsymbol{a}_{\mathrm{mi}}^{\mathrm{T}} \boldsymbol{s}_i(n) + v_m(n), \quad m = 1, 2 \tag{1}$$

where n is the time index, $\boldsymbol{a}_{\mathrm{mi}} = [a_{\mathrm{mi}}(0), \ldots, a_{\mathrm{mi}}(L_a - 1)]^{\mathrm{T}}$ is the L_a-length room impulse response between the ith source and the mth microphone, $\boldsymbol{s}_i(n) = [s_i(n), \ldots, s_i(n - L_a + 1)]^{\mathrm{T}}$ is the ith source signal vector and $v_m(n)$ is the uncorrelated environment noise at the mth microphone. For each impulse response $\boldsymbol{a}_{\mathrm{mi}}$, the location of the highest peak, n_{mi}, denotes the arrival time of the ith source at the mth microphone. The TDOA of the ith source with respect to two microphones is defined as

$$\tau_i = \frac{n_{2i} - n_{1i}}{f_s} \tag{2}$$

where f_s denotes the sampling rate. TDOA is a key parameter in sound source localization, since the DOA can be calculated directly from the TDOA using $\tau_i = \frac{d\cos(\theta_i)}{c}$, where c denotes the speed of sound.

The goal is to estimate the number of sources, N, as well as their TDOAs $\{\tau_1, \ldots, \tau_N\}$ from the microphone signals. The main challenges for source counting and localization are environment noise, the presence of multiple sources and reverberation. In addition to this, spatial aliasing can be introduced when the two microphones are far apart. To address these challenges, we propose a joint source counting and localization framework, based on T-F IPD, as described below.

IV. TDOA Estimation

The proposed joint source counting and localization framework consists of three main blocks, namely IPD calculation, T-F weighting and ICR (see Fig. 1). The first two blocks will be introduced in this section, while the third block will be addressed in Section V.

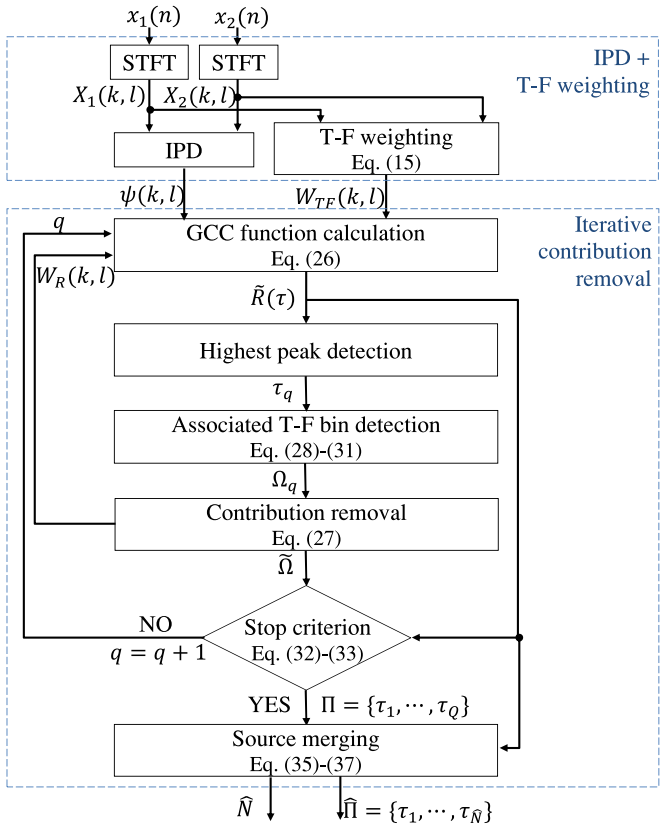


Fig. 1. Block diagram of the proposed joint source counting and localization method. Input: microphone signals x_1 and x_2. We initialize $q = 1$ and $W_R(k,l) = 1 \ \ \forall k, l$. Output: the number of sources $\hat{N}$ and TDOAs $\{\tau_1, \ldots, \tau_{\hat{N}}\}$.

A. IPD-Based TDOA Estimation

We first derive the framework for TDOA estimation based on T-F IPD in anechoic and noise-free environments and then extend it to noisy and reverberant scenarios.

In the anechoic and noise-free scenario, the signal received at the mth microphone can be simplified as

$$x_m(n) = \sum_{i=1}^{N} a_{\mathrm{mi}} s_i(n - n_{\mathrm{mi}}) \tag{3}$$

where n_{mi} and a_{mi} are the transmitting delay and attenuation from s_i to the mth microphone, respectively. Transforming the microphone signals into the T-F domain using the short-time Fourier transform (STFT), we can rewrite the microphone signal for each T-F bin as

$$X_m(k,l) = \sum_{i=1}^{N} a_{\mathrm{mi}} S_i(k,l) \mathrm{e}^{j2\pi f_k n_{\mathrm{mi}}/f_s} \tag{4}$$

where k and l are the frequency and frame indices, respectively, f_k represents the frequency at the k-th frequency bin, and S_i is the STFT of s_i.

Assuming that only one source s_i is active, the IPD between two microphones can be expressed as

$$\psi_{\mathrm{s}}(k,l) = \angle \frac{X_2(k,l)}{X_1(k,l)} = 2\pi f_k \tau_i + 2\pi p_k \tag{5}$$

where τ_i is the TDOA of the source, the wrapping factor p_k is a frequency-dependent integer and $2\pi p_k$ represents possible phase wrapping, and $\psi_{\rm s}$ is constrained to be in the range $[-\pi, \pi]$ after mod(2π) operation. If the phase wrapping part can be neglected, the IPD $\psi_{\rm s} = 2\pi f_k \tau_i$ in (5) will vary linearly with respect to the frequency, f_k, with a variation slope being τ_i. We call this linear variation a phase variation line (PVL) of τ_i. The wrapping factor

$$p_k = |2\pi f_k \tau_i|_{2\pi} \tag{6}$$

is an integer determined jointly by the TDOA τ_i and the frequency f_k, where $|\cdot|_{2\pi}$ retains the integer after mod(2π) operation. The larger the inter-microphone distance and the higher the frequency, the more wrapping is expected. This phenomenon is called *spatial aliasing ambiguity*. In theory, when d is smaller than half the wavelength, no phase wrapping occurs, i.e., $2\pi p \equiv 0$.

When N sources are active, we assume the audio mixtures comply with W-disjoint orthogonality, meaning that in each T-F bin at most one source is dominant [37]. In this case, the IPD between two microphones can be expressed as

$$\psi_{\rm m}(k,l) = \angle \frac{X_2(k,l)}{X_1(k,l)} = 2\pi f_k \tau_{\rm kl} + 2\pi p_{\rm kl} \tag{7}$$

where $\tau_{\rm kl}$ and $p_{\rm kl}$ denote TDOA and phase unwrapping of the dominant source in the (k,l)-th bin, respectively, $\tau_{\rm kl} \in \{\tau_1, \ldots, \tau_N\}$, and $\psi_{\rm m}$ is constrained to be in the range $[-\pi, \pi]$ after mod(2π) operation.

When no phase wrapping happens (i.e., $p_{\rm kl} \equiv 0$), a clustering algorithm can be applied to IPD to estimate both the number of sources and their TDOAs [12]. However, for larger inter-microphone distances with severe phase wrapping (i.e., $|p_{\rm kl}| > 1$), the clustering algorithm will fail due to unassociated frequency-dependent wrapping factors with different sources.

The phase wrapping ambiguity is mainly caused by the extra term $2\pi p_{\rm kl}$. Since $\mathrm{e}^{j(2\pi f_k \tau_{\rm kl} + 2\pi p_{\rm kl})} = \mathrm{e}^{j2\pi f_k \tau_{\rm kl}}$, we propose a new framework which works in the exponential domain to avoid this ambiguity. Instead of estimating the TDOAs directly from the IPD, the framework employs an exhaustive search in the TDOA domain with the cost function defined as

$$R(\tau) = \left| \sum_{k,l} \mathrm{e}^{j\psi_{\rm m}(k,l)} \mathrm{e}^{-j2\pi f_k \tau} \right| = \left| \sum_{k,l} \mathrm{e}^{j2\pi f_k (\tau_{\rm kl} - \tau)} \right|. \tag{8}$$

As shown in (8), the wrapping term $2\pi p_{\rm kl}$ disappears due to the exponential operation. Assuming W-disjoint orthogonality with each source i exclusively occupying one set of T-F bins $\mathbb{B}_i$, the cost function can be further written as

$$R(\tau) = \left| \sum_{i=1}^{N} \sum_{k,l \in \mathbb{B}_i} \mathrm{e}^{j2\pi f_k (\tau_i - \tau)} \right|. \tag{9}$$

From (9), $R(\tau)$ tends to show a peak value at $\tau = \tau_i$. Therefore, (9) can be approximated as a sum of N peaks which originate from the N sources. This is expressed as

$$R(\tau) \approx \sum_{i=1}^{N} B_i \delta(\tau - \tau_i) \tag{10}$$

where B_i is the number of T-F bins in the set $\mathbb{B}_i$, which is practically unknown. The TDOAs and number of sources can thus be detected from the peaks of $R(\tau)$.

The IPD-based algorithm is essentially equivalent to the well-known GCC-PHAT algorithm [18], whose cost function to be maximized is defined as

$$\begin{aligned} R_{\rm GCC}(\tau) &= \left| \sum_{l} \sum_{k} \frac{X_1^*(k,l) X_2(k,l)}{|X_1(k,l) X_2(k,l)|} \mathrm{e}^{-j2\pi f_k \tau} \right| \\ &= \left| \sum_{l} \sum_{k} \mathrm{e}^{j\psi_{\rm m}(k,l)} \mathrm{e}^{-j2\pi f_k \tau} \right| = R(\tau) \end{aligned} \tag{11}$$

and the TDOA is estimated as

$$\tau_o = \arg\max_{\tau} R_{\rm GCC}(\tau). \tag{12}$$

As shown in (11), the GCC-PHAT algorithm and the proposed IPD-based algorithm have the same cost function. However, the two algorithms are derived from different perspectives. Assuming a single source, GCC-PHAT maximizes the correlation between two microphone signals and introduces phase weighting to improve the robustness to reverberation. In contrast, the proposed algorithm is derived based on the concept of IPD between two microphone signals and does not require the single-source assumption. This provides a theoretical grounding for multi-TDOA estimation. Combining IPD with subsequent T-F weighting and ICR leads to a solution for multi-source counting and localization. For simplicity, we refer to the cost functions in both (8) and (11) as the GCC function.

B. T-F Weighting

The IPD-based algorithm was derived based on the assumption of anechoic, noise-free and W-disjoint orthogonality conditions. These assumptions are rarely met in practice, thus leading to degraded performance in TDOA estimation and source counting. We use T-F processing to exploit the nonstationarity and sparsity of audio signals to address the challenge of ambient noise and overlap of multiple sources. We employ two T-F weighting schemes, namely SNR weighting and coherence weighting [12], [38], [41], [45]. In this case, the T-F weighted GCC function becomes

$$R_{\rm w}(\tau) = \left| \sum_{k,l} W_{\rm TF}(k,l) \frac{X_1^*(k,l) X_2(k,l)}{|X_1(k,l) X_2(k,l)|} \mathrm{e}^{-j2\pi f_k \tau} \right| \tag{13}$$

with

$$W_{\rm TF}(k,l) = W_{\rm SNR}(k,l) W_{\rm coh}(k,l) \tag{14}$$

being the product of SNR weight and coherence weight.

We use SNR weighting to improve robustness to ambient noise. This is performed based on the SNR at an individual T-F bin, namely local SNR. T-F bins with high local SNRs are less

affected by ambient noise and thus are given higher weights in the GCC function [12]. The local SNR $\lambda(k,l)$ is calculated as

$$\lambda(k,l) = \min\left(\frac{P_{x_1}(k,l)}{P_{v_1}(k,l)} - 1, \frac{P_{x_2}(k,l)}{P_{v_2}(k,l)} - 1\right) \tag{15}$$

where $P_{x_m}(k,l) = |X_m(k,l)|^2, m = 1,2$, is the power of the mth microphone signal, while $P_{v_m}(k,l)$ is the power of the noise signal. Assuming an ideal case where the noise is stationary and the first L_v frames of the microphone signal contain only noise, $P_{v_m}(k,l)$ is time-invariant and can be calculated as

$$P_{v_m}(k) = \frac{1}{L_v}\sum_{l=1}^{L_v}|X_m(k,l)|^2. \tag{16}$$

To determine the SNR weight we use

$$W_{\text{snr}}(k,l) = \begin{cases} 1 & \lambda(k,l) > \lambda_{\text{TH}} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

where λ_{TH} is a predefined threshold.

To reduce the influence of overlapped sources on the GCC function, a coherence weighting scheme is employed to detect and discard the T-F bins with multiple active sources [41]. The coherence at the (k,l)th bin is defined as

$$r(k,l) = \left|\frac{\text{E}(X_1(k,l)X_2^*(k,l))}{\sqrt{\text{E}(X_1(k,l)X_1^*(k,l))}\sqrt{\text{E}(X_2(k,l)X_2^*(k,l))}}\right| \tag{18}$$

where the expectation $\text{E}(\cdot)$ is approximated by averaging among $2C+1$ consecutive time frames. For instance,

$$\text{E}(X_1(k,l)X_2^*(k,l)) = \frac{1}{2C+1}\sum_{l'=l-C}^{l+C}X_1(k,l')X_2^*(k,l'). \tag{19}$$

Based on the continuity of speech signals along time, a T-F bin is believed to be one-source active if its coherence is higher than a threshold, i.e.,

$$W_{\text{coh}}(k,l) = \begin{cases} 1 & r(k,l) > r_{\text{TH}} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where r_{TH} is a predefined threshold.

The choice of λ_{TH} and r_{TH} determines the number of T-F bins that can be reliably employed for the subsequent source counting and localization, and hence is crucial to the performance of the whole system in noisy environments. A discussion about the choice of these parameters will be given in Section V-D.

V. Joint Source Counting and Localization

After T-F weighting, the next step is to count and localize the sources.

Ideally, the GCC function will be a sequences of peaks (as in (10)), whose number is equal to the number of sources. However, the interaction between multiple sources and multiple paths leads to ambiguities in the interpretation of the peaks of the GCC function, thus making it difficult to get the number of sources and their TDOAs. As an example, Fig. 2 depicts the GCC function for four sources, whose locations are indicated with red circles. The microphone signal is simulated using the method in Section VI-C, with an inter-microphone distance of 1 m and a direct to reverberation ratio (DRR) of 0 dB. The GCC function contains more peaks than sources and it is difficult to distinguish a peak associated with a true source from a spurious one by solely observing the GCC function.

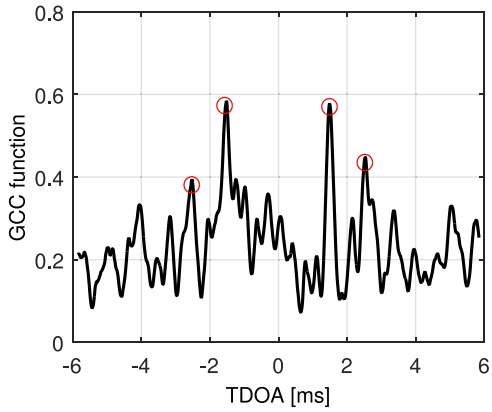


Fig. 2. Peak ambiguities between multiple sources and multiple paths for four sources. The locations of the four sources are indicated with red circles. The inter-microphone distance is 1 m, DRR = 0 dB.

To demonstrate the peak ambiguity problem, we employ two simplistic acoustic systems. The first system consists of two microphones and two sources (s_1 and s_2) in an anechoic scenario, while the second system consists of two microphones and one source (s_1) in a reverberant scenario where only the first reflection is considered. The transfer functions of the two systems are, respectively, expressed as

$$\begin{cases} x_1^{\text{I}}(n) = a_{11}^{\text{I}}s_1(n - t_{11}^{\text{I}}) + a_{12}^{\text{I}}s_2(n - t_{12}^{\text{I}}) \\ x_2^{\text{I}}(n) = a_{21}^{\text{I}}s_1(n - t_{21}^{\text{I}}) + a_{22}^{\text{I}}s_2(n - t_{22}^{\text{I}}) \end{cases} \tag{21}$$

and

$$\begin{cases} x_1^{\text{II}}(n) = a_{11}^{\text{II}}s_1(n - t_{11}^{\text{II}}) + a_{12}^{\text{II}}s_1(n - t_{12}^{\text{II}}) \\ x_2^{\text{II}}(n) = a_{21}^{\text{II}}s_1(n - t_{21}^{\text{II}}) + a_{22}^{\text{II}}s_1(n - t_{22}^{\text{II}}) \end{cases} \tag{22}$$

where a_{11}^{I}, a_{12}^{I}, a_{21}^{I}, a_{22}^{I}, a_{11}^{II}, a_{12}^{II}, a_{21}^{II}, a_{22}^{II} represent the attenuation coefficients while t_{11}^{I}, t_{12}^{I}, t_{21}^{I}, t_{22}^{I}, t_{11}^{II}, t_{12}^{II}, t_{21}^{II}, t_{22}^{II} represent the transfer delays.

In the second system, s_2 is replaced by a reflection of s_1. We use the same attenuation coefficients and transfer delays for the two systems by arbitrarily setting $a_{11} = 1$, $a_{12} = 0.4$, $a_{21} = 1$, $a_{22} = 0.4$, and $t_{11} = 0$, $t_{12} = 4$, $t_{21} = 1$, $t_{22} = 7$ (the superscript $(\cdot)^{\text{I}}$ and $(\cdot)^{\text{II}}$ are neglected for clarity). For a sampling rate of 8 kHz, the TDOAs in the first (two-source) system are $\tau_1^{\text{I}} = -0.125$ ms and $\tau_2^{\text{I}} = -0.375$ ms; the TDOA in the second (one-source) system is $\tau_1^{\text{II}} = -0.125$ ms. We use 10 s long male and female speech files for the two sources.

Fig. 3 shows the IPDs and GCCs of the two systems. The GCC plots present multiple peaks. Although the true TDOAs are contained in these peaks, it is difficult to tell which one is the true value. However, the TDOA corresponding to the highest peak is always a true one. In contrast to the ambiguous peaks in the GCC plots, the IPD plots show a clear difference. In Fig. 3(a)

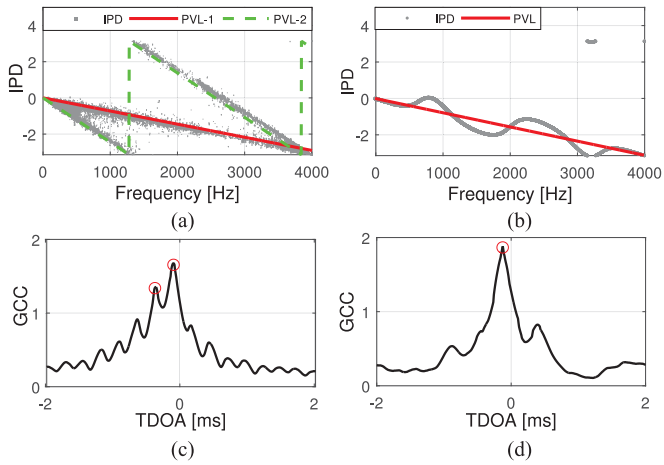


Fig. 3. IPD and GCC of an anechoic two-source system and a reverberant one-source system: (a) IPD and (c) GCC of the two-source system; (b) IPD and (d) GCC of the one-source system.

the IPD of the first (anechoic and two-source) system can be easily fitted with two PVLs (one line $2\pi f\tau_1^{\mathrm{I}}$ and one phase-wrapped line $2\pi f\tau_2^{\mathrm{I}}$). In Fig. 3(b) for a reverberant source, only one curve can be observed, which fluctuates vigorously along the PVL line of the true TDOA ($2\pi f\tau_1^{\mathrm{II}}$).

Exploiting the discrimination ability of the IPD plot, we propose an ICR algorithm, as shown in Fig. 1, to count and localize multiple sound sources from the IPD and GCC plots. The basic idea is to detect a source from the highest peak of the GCC plot, and from the IPD plot to detect the T-F bins that are associated with this source. The detected T-F bins are subsequently removed from the IPD plot so that a new GCC function can be calculated to detect the next source. In this way, all sources can be detected iteratively. When designing the algorithm, several challenges arise: how to detect the T-F bins that are associated with a target source; how to remove them from the IPD plot; and how to stop the iteration when all the sources are detected. These issues will be addressed below.

A. Contribution Removal

In the anechoic scenario, the detection and removal of the T-F bins associated with a source can be easily conducted since the IPD of the source fits well with the PVL of its TDOA. Suppose that the source q is detected as the highest peak of the GCC plot with its TDOA τ_q using (12). The correlation between a T-F ((k,l)th) bin with the source can be measured by the distance between the IPD $\psi(k,l) = \angle\frac{X_2(k,l)}{X_1(k,l)}$ and the PVL of the source. The distance is expressed as

$$\rho(k,l,\tau_q) = \left|\angle \mathrm{e}^{j(\psi(k,l)-2\pi f_k\tau_q)}\right| \tag{23}$$

where the exponential operation can cancel out the phase wrapping ambiguity. We assume this T-F bin belongs to the source if the distance is sufficiently small, i.e.,

$$(k,l)\in\Omega_q \quad \forall\rho(k,l,\tau_q) < \rho_{\mathrm{TH}} \tag{24}$$

where Ω_q denotes the T-F set that associates with the source and ρ_{TH} is a threshold. The removal of the detected bins can be realized by applying another T-F weight to the GCC function (13), i.e.,

$$\tilde{R}(\tau) = \left|\sum_{k,l} W_{\mathrm{R}}(k,l)W_{\mathrm{TF}}(k,l)\mathrm{e}^{j\psi(k,l)}\mathrm{e}^{-j2\pi f_k\tau}\right| \tag{25}$$

where W_{TF} is defined in (14) and W_{R} denotes the weight for contribution removal, which is calculated by

$$W_{\mathrm{R}}(k,l) = 0 \quad \forall(k,l)\in\Omega_q. \tag{26}$$

In the reverberant scenario, the detection and removal of the T-F bins associated with a source becomes more complicated, because the IPD of the source does not well fit the PVL of its TDOA. For instance, the IPD of a reverberant source in Fig. 3(b) spans a wider space than an anechoic source in Fig. 3(a). In this case, it is difficult to detect all the T-F bins that belong to the target source. After applying the removal procedure which is designed for the anechoic scenario, residual T-F bins (of the target source) still exist and will affect the iteration of the next round.

To solve this problem, we propose an improved detection and removal method. In Fig. 3(b) the IPD of a reverberant source is fluctuating around the true PVL. Based on this observation, the distance between the (k,l)th bin with the PVL line is modified as the distance between the bin and a set of parallel lines, which is defined as

$$\rho'(k,l,\tau_q) = \left|\angle \mathrm{e}^{j(\psi(k,l)-(2\pi f_k\tau_q+\delta'_q))}\right| \tag{27}$$

where δ'_q denotes the optimal shift (along the IPD axis) from the original PVL. The optimal parallel line is selected from the set of parallel lines which can capture the largest number of T-F bins. This is expressed as

$$\delta_q = \arg\min_{\delta}\sum_{k,l}\left|\angle \mathrm{e}^{j(\psi(k,l)-(2\pi f_k\tau_q+\delta))}\right| \tag{28}$$

and

$$\delta'_q = \pm\delta_q. \tag{29}$$

As indicated in (29), two parallel lines, lying above and below the PVL line, respectively, are used. The optimization problem in (28) is solved by using an exhaustive search in the range $[-\pi/3,\ \pi/3]$. Similarly to (24), the association between the (k,l)th T-F bin and the qth source can be determined by

$$(k,l)\in\Omega_q \quad \forall\rho'(k,l,\tau_q) < \rho_{\mathrm{TH}}, \tag{30}$$

where ρ_{TH} is a predefined threshold (see the discussion in Section V-D). The removal is performed by (25).

B. Stop Criterion

When performing contribution removal iteratively, we employ a stop criterion so that the number of sources can be reliably counted. We note that the GCC function is sparse with strong peaks when one or several sources are active, and becomes noisy, with no evident peaks, when the contribution from the sources has been mostly removed. Thus, the stop criterion is mainly based on the sparsity of the GCC function, which can

be measured by the kurtosis value [53]. In addition to this, the iteration will stop when all the bins are removed. In summary, the iteration stops if it reaches a predefined maximum number $Q_{\max}$; or if, after contribution removal, the number of the remaining bins is small enough, i.e.,

$$\text{If} \quad \text{size}\{\tilde{\Omega}\} < 0.01 \cdot \text{size}\{\Omega\}, \quad \text{STOP = TRUE} \tag{31}$$

where $\tilde{\Omega}$ denotes the complement of the set $\Omega = \{\Omega_1, \ldots, \Omega_q\}$; or if no evident peak is detected in the GCC function. The GCC function has no evident peak present if its kurtosis value is sufficiently small, i.e.,

$$\text{If} \quad \text{kurt}(\tilde{R}) < K_{\text{TH}}, \quad \text{STOP = TRUE} \tag{32}$$

where $\text{kurt}(\cdot)$ denotes the kurtosis of the argument, the GCC function $\tilde{R}$ is given by (25), and K_{TH} is a predefined threshold (see the discussion in Section V-D). We set $Q_{\max} = 10$ since we observed that after ten iterations the residual T-F bins usually do not provide reliable TDOA information and, moreover, by introducing other stop criteria, the algorithm usually terminates before ten iterations.

After iteration, we obtain Q TDOAs initially denoted as

$$\Pi = \{\tau_1, \ldots, \tau_Q\}. \tag{33}$$

C. Source Merging

An advantage of the proposed ICR algorithm is its ability to detect and remove the residual T-F bins that are not removed in an iteration. When the iteration is completed, a source could be repeatedly detected during the iteration. Thus we use a postprocessing scheme to merge closely located sources, based on their distance and the strength of the peaks.

The distance criterion is expressed as

$$\text{If} \quad |\tau_p - \tau_q| < \frac{d \sin(A_{\min})}{c}, \quad \tau_m \leftarrow \{\tau_p, \tau_q\} \tag{34}$$

where $A_{\min}$ is a minimum separation angle, τ_p and τ_q are two detected TDOAs in Π, and $\tau_m \leftarrow \{\tau_p, \tau_q\}$ denotes the merge of the two, which can be implemented as

$$\tau_m = \begin{cases} \tau_p & \tilde{R}_o(\tau_p) > \tilde{R}_o(\tau_q) \\ \tau_q & \text{otherwise} \end{cases} \tag{35}$$

where $\tilde{R}_o$ denotes the original GCC function by (25) in the first iteration. We observed that the correct estimation usually presents the highest GCC value among all the closely located candidates and thus in (35) we use the estimate with highest GCC value as the location of the merged source.

The criterion on the strength of the GCC peak of a detected source is expressed as

$$\text{If} \quad \tilde{R}_o(\tau_q) < R_{\text{TH}}, \quad \Pi \leftarrow \Pi \backslash \tau_q \tag{36}$$

where the threshold R_{TH} is set as the median value of $\tilde{R}_o$.

After postprocessing, we obtain $\hat{N}$ TDOAs denoted as

$$\hat{\Pi} = \{\tau_1, \ldots, \tau_{\hat{N}}\}. \tag{37}$$

TABLE II
PARAMETERS USED BY THE PROPOSED ICR ALGORITHM

Parameter	Equation	Value
λ_{TH}	(17)	5 dB
C	(19)	2
r_{TH}	(20)	0.9
ρ_{TH}	(30)	0.3
K_{TH}	(32)	3
$A_{\min}$	(34)	10°

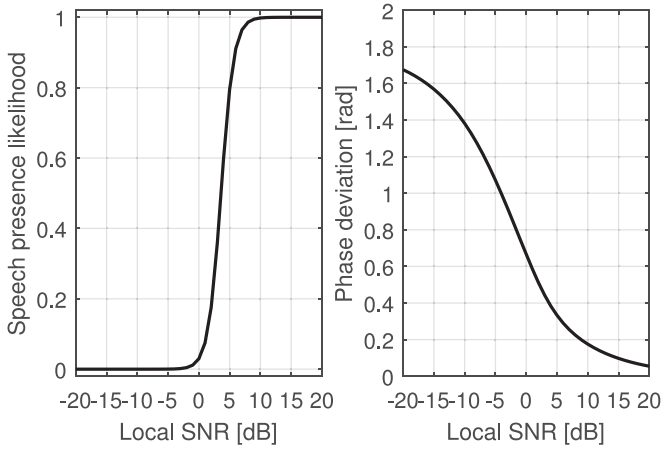


Fig. 4. (a) Speech-presence likelihood and (b) averaged phase deviation versus local SNR.

D. Parameters

The parameters used by the proposed algorithm are summarized in Table II. Under the sampling rate of 8 kHz, we choose STFT window length of 1024 with an overlap size of 512. When calculating the GCC function (25), we set the search area as $[-\frac{d}{c}, \frac{d}{c}]$, with the searching step 10^{-5} s. The selection of the parameters is justified below.

Regarding coherence weighting in (18)–(20), we choose the parameters by referring to [27], [41], [45]. We calculate the coherence over 5 ($C = 2$) consecutive frames and use the coherence threshold $r_{\text{TH}} = 0.9$ for one-source dominance detection. Regarding SNR weighting in (15)–(17), we determine the threshold based on the speech-presence likelihood $p_{\text{H}}(k, l)$, which can be modelled as a function of local SNR $\lambda(k, l)$ as [52]

$$p_{\text{H}}(k, l) = \left(1 + (1 + \xi)\text{e}^{-\frac{\xi(1+\lambda(k,l))}{1+\xi})}\right)^{-1} \tag{38}$$

where $\xi = 15$ dB is the a priori SNR. Fig. 4(a) depicts the variation of the speech-presence likelihood with respect to the local SNR. We choose the SNR threshold $\lambda_{\text{TH}} = 5$ dB so that the speech-presence likelihood is close to 0.8.

Regarding the distance threshold in (30), we aim to capture the most T-F bins that are associated with a target source with the smallest distance. For this aim, we investigate how an additive noise affects the phase of the source signal with a simple simulation. We use 200 000 samples of complex-valued source signals plus complex-valued noise signals at different (local)

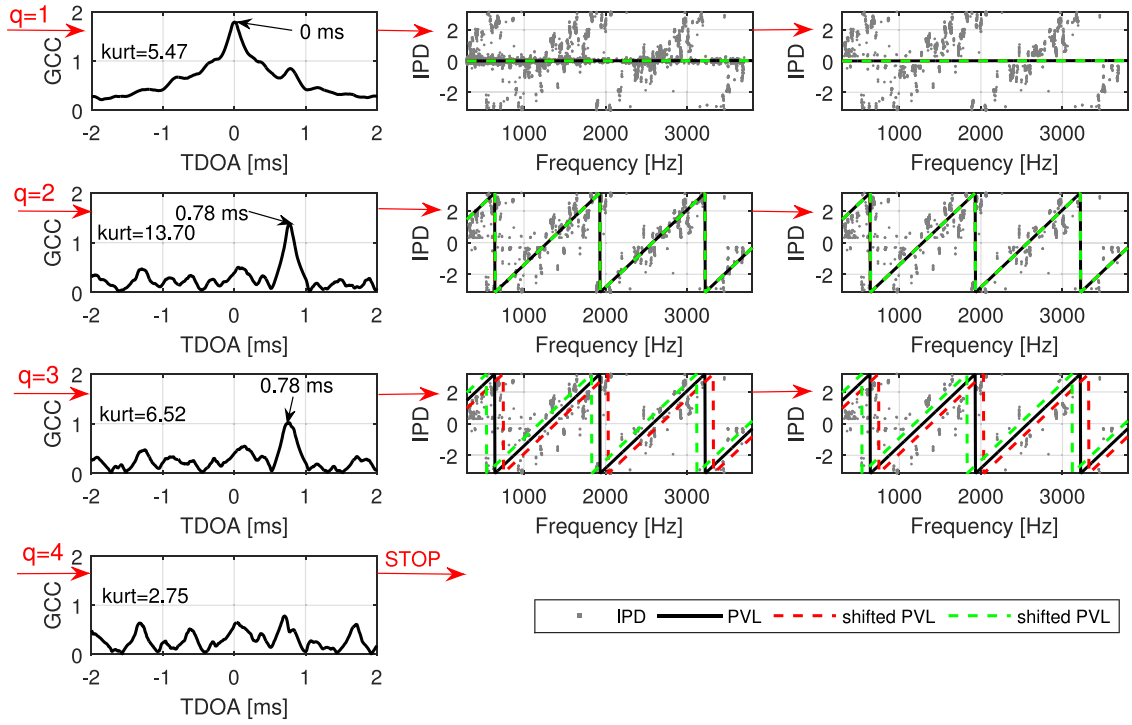


Fig. 5. Intermediate results when applying the ICR algorithm to the two-source scenario (simulated by the image-source method with a reverberation time of 400 ms). Each row depicts the results in one (qth) iteration.

SNRs. The real and imaginary parts of the source signal are both independent Gaussian processes (with mean and variance being 0 and 1, respectively). The real and imaginary parts of the noise signal are also independent Gaussian processes, but with variable amplitudes for different SNRs. For each SNR, we calculate the standard phase deviation of all the samples. Fig. 4(b) depicts how the phase of the source signal deviates at different SNRs. In Fig. 4(b) the phase deviation is around 0.3 at the SNR of 5 dB. We thus choose the distance threshold as $\rho_{\mathrm{TH}} = 0.3$, in correspondence to $\lambda_{\mathrm{TH}} = 5$ dB.

The kurtosis value in (32) is an important measure to judge whether the iteration can stop. The GCC function shows peaks when the sources are active, and becomes noisy when no source is active. Since the kurtosis of a Gaussian noise is around 3, we choose $K_{\mathrm{TH}} = 3$ as the stop threshold.

The minimum separation angle is a user-defined threshold that determines the resolution of the TDOA estimation. We choose $A_{\mathrm{min}} = 10^{\circ}$ as suggested in [38].

E. Example

We show an example of applying the proposed ICR algorithm to a two-source scenario simulated using the image-source method [50] in an enclosure of size 7 m $\times$ 9 m $\times$ 3 m with a reverberation time of 400 ms. The two microphones are 0.3 m apart; the two sources are placed 2 m away at 90° and 30°, with the TDOAs being 0 and 0.76 ms, respectively. Fig. 5 depicts the intermediate processing results by the proposed ICR algorithm (see Fig. 1).

In Fig. 5, each row depicts one (the qth) iteration; the first column depicts the calculated GCC function and the detected highest peak; the second column depicts the PVL of the qth source and its shifted PVL, as well as the IPD; the third column depicts the IPD after removing the T-F bins associated with the qth source. The kurtosis of the GCC function in each iteration is also given in the first column.

Due to multiple reflections, the IPDs of the T-F bins associated with each source vary vigorously and irregularly, but still around the PVL of the source (see Fig. 5). In the first iteration, the peak from the first source is dominant in the GCC function. After removing the T-F bins associated with the first source, the peak of the second source becomes dominant in the GCC function. The kurtosis value of the GCC function for $q = 2$ is even higher than the one for $q = 1$. The second and third iterations remove the T-F bins associated with the second source. The utility of the shifted PVL can be clearly seen in the third iteration, where the shifted PVL can capture the residual bins that are not captured by the original PVL. When the contribution of the second source is removed gradually, the GCC function becomes noisy and the kurtosis value becomes smaller. The iteration terminates at $q = 4$ since the kurtosis value of the GCC function is smaller than 3. We obtain three TDOAs: [0, 0.78, 0.78] ms, which are merged into two estimates: 0 and 0.78 ms.

VI. Experimental Results

A. Algorithms for Comparison

We compare the proposed algorithm (ICR) with another two source counting algorithms: direct peak counting (DC) and DEMIX. DC counts the number of sources based on the peaks of the GCC function. Some principles, which are presented in [38] for source counting based on a DOA histogram, can also be

employed for this task, namely the distance between two sources should be larger than 10° and the peak of a source should be higher than a threshold, which is defined as a function of previously detected peaks (cf., (14)–(16) in [38]). DEMIX uses a clustering algorithm applied to signal amplitude for source counting [11]. After clustering, the source localization in each cluster can be calculated with a GCC-like function. We use the source code provided by Arberet *et al.* [11].

The comparison is performed in acoustic scenarios simulated with artificially generated room impulse responses (Sections VI-C and VI-D), image-source method based room impulse responses (Section VI-E), and real-recorded acoustic impulse responses (Section VI-F).

B. Evaluation Measures

Since there are multiple sources and multiple estimates, it is difficult to associate each estimate with a correct value for calculating the estimation error. We thus evaluate the localization performance under two aspects. First, we count the number of correctly detected sources and evaluate the source counting performance in terms of recall rate, precision rate and F-score. We assume that a TDOA estimation is correct if its corresponding DOA is close enough to a true source (i.e., the DOA difference is smaller than 10°, the minimum separation angle in (34)). Second, for the correctly detected sources, we evaluate the localization accuracy with the TDOA estimation error. These measures are defined as below.

Recall rate and precision rate evaluate the performance in terms of miss-detections and false alarms, respectively, while F-score evaluates the source counting performance globally. Suppose the true number of sources is N, and the estimated number of sources is $\hat{N}$ with the number of correct ones being $\hat{N}_c$. The three measure are, respectively, defined as

$$R_{\text{rate}} = \frac{\hat{N}_c}{N}, \quad P_{\text{rate}} = \frac{\hat{N}_c}{\hat{N}}, \quad F_{\text{score}} = 2\frac{P_{\text{rate}} \cdot R_{\text{rate}}}{P_{\text{rate}} + R_{\text{rate}}}. \quad (39)$$

The global measure F-score can be interpreted as the harmonic average of the precision and recall, reaching its best value at 1 and worst at 0.

For each correctly detected source, the TDOA estimation error is defined as

$$\tau_d = |\tau_o - \tau_e| \quad (40)$$

where τ_o and τ_e denote the true and estimated TDOAs, respectively.

C. Simulation Environment for Artificial Impulse Response

Four inter-microphone distances are used: $\{0.3, 1, 3, 6\}$ m. For each inter-microphone distance, seven source directions from 0° to 180°, with an interval of 30°, are considered. Speech files (six males and six females) are used for the experiment, each 10 s long and sampling rate 8 kHz.

The impulse response between the jth source and the ith microphone is modelled as

$$h_{ij}(n) = h_{ij}^{\text{d}}(n) + h_{ij}^{\text{r}}(n) \quad (41)$$

TABLE III
NUMBER OF SOURCES VERSUS DOA IN THE SIMULATION

Number of Sources	DOA [$^\circ$]
2	60, 120
3	60, 90, 120
4	30, 60, 120, 150
5	30, 60, 90, 120, 150
6	0, 30, 60, 120, 150, 180

where h^{d} and h^{r} denote the direct and reverberant part, respectively [11]. The direct part is modelled as a delayed impulse as

$$h_{ij}^{\text{d}}(n) = \delta(n - n_{ij}) \quad (42)$$

with

$$\begin{cases} n_{1j} = n_0 \\ n_{2j} = n_0 + \dfrac{d\cos(\theta_j)}{c} f_s \end{cases}, \quad (43)$$

where $n_0 = 100$ ms denotes a constant reference time point for all the sources. The reverberant part is modelled as an independent Gaussian noise process $h_{ij}^{\text{r}}(n) \sim \mathcal{N}(0, \sigma^2(n - n_{ij} - n_1))$ with

$$\sigma^2(n) = \begin{cases} 10^{-\alpha n}\sigma_R^2 & 0 < n - (n_{ij} + n_1) < T_r f_s \\ 0 & \text{otherwise} \end{cases} \quad (44)$$

with $\alpha = 6/T_r$, so as to have an exponential decrease of -60 dB at the end of the reverberation part, and with $n_1 = 20$ ms being the distance between the direct and the reverberant part and $T_r = 150$ ms being the length of the reverberation. The parameter σ_R^2 controls the DRR which is defined as

$$\text{DRR} = 10\log_{10}\frac{\sum_n (h^{\text{d}}(n))^2}{\sum_n (h^{\text{r}}(n))^2}. \quad (45)$$

In this way, all the sound sources are modelled as plane waves with the reverberation density (DRR) controlled by σ_R^2. We consider seven different DRRs increasing from -10 to 20 dB, with an interval of 5 dB.

The number of sources varies from 2 to 6. The directions of the sources are selected based on the number of sources. Table III lists the relationship between the two terms. For each geometrical configuration we implement 15 instances. In each instance, the speech is randomly selected from the 12 files while the reverberant part h^{r} of the impulse response is generated independently. The microphone signals are generated via convolution between the speech files and the corresponding impulse responses.

Speech-shaped Gaussian noise, computed by filtering Gaussian noise through an FIR filter whose frequency response matches the long-term spectrum of speech [51], is added at different SNRs (from -10 to 30 dB).

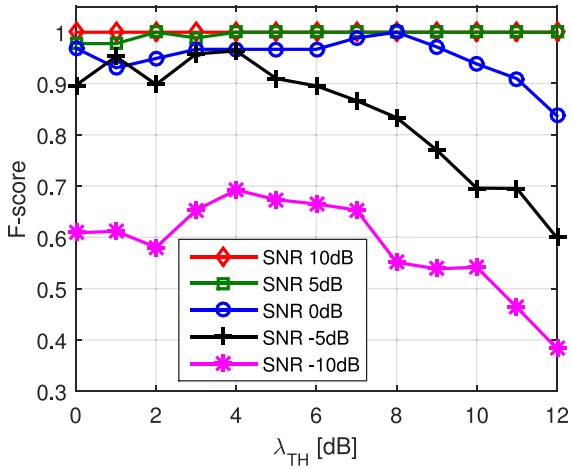


Fig. 6. Performance (F-score) of the ICR algorithm versus λ_{TH} for different SNRs from -10 to 10 dB. The inter-microphone distance is 1 m, 4 sources, DRR $= 20$ dB, $\rho_{\mathrm{TH}} = 0.3$.

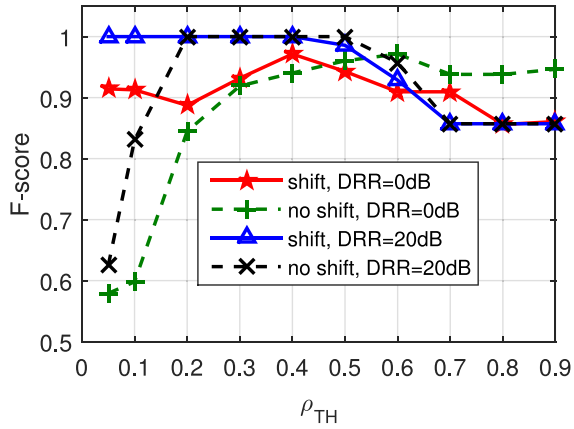


Fig. 7. Performance (F-score) of the ICR algorithm versus ρ_{TH} for two DRRs, 0 and 20 dB, respectively. The inter-microphone distance is 1 m, 4 sources, SNR $= 30$ dB, $\lambda_{\mathrm{TH}} = 5$ dB.

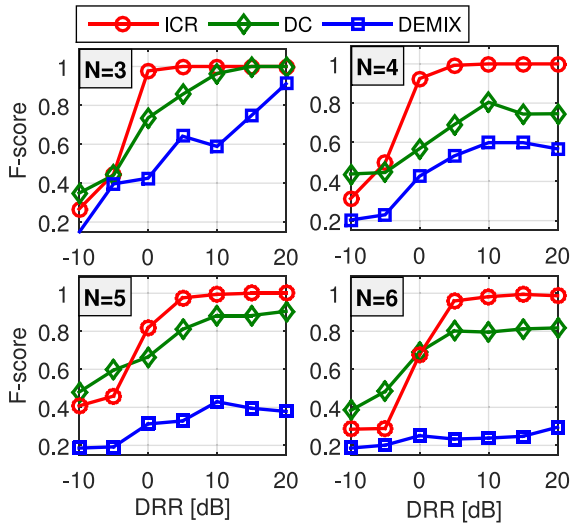


Fig. 8. Performance (F-score) comparison of the source counting algorithms for different numbers of sources N and DRRs. The inter-microphone distance is 1 m, SNR $= 30$ dB.

D. Results From Artificial Room Impulse Response

In this experiment we first examine how the performance of the proposed ICR algorithm varies with the two parameters λ_{TH} and ρ_{TH}. Then we compare the performance of the three algorithms in conditions with varying numbers of sources N, inter-microphone distances d, reverberation densities (DRR) and noise intensities (SNR).

1) Influence of λ_{TH}: Under the condition $d = 1$ m, $N = 4$, DRR $= 20$ dB, SNRs increasing from -10 to 10 dB, with an interval of 5 dB, and $\rho_{\mathrm{TH}} = 0.3$, we examine how the performance of the ICR algorithm varies when λ_{TH} increases from 0 to 12 dB, with an interval of 1 dB. Fig. 6 depicts the F-scores obtained by the ICR algorithm. The performance of the ICR algorithm degrades with the increase of the noise level. In high SNRs (5 and 10 dB), the F-score keeps almost constant for various λ_{TH}. For low SNRs (-10, -5 and 0 dB) and $\lambda_{\mathrm{TH}} \geq 2$ dB, the F-score tends to rise with increasing λ_{TH} until reaching a peak value, and then drops quickly with increasing λ_{TH}. For SNRs -5 and -10 dB the peak is reached when $\lambda_{\mathrm{TH}} = 4$ dB, while for SNR 0 dB the peak is reached when $\lambda_{\mathrm{TH}} = 8$ dB. The observations demonstrate that T-F weighting can improve the performance of the ICR algorithm in noisy environments. The observations confirm our choice $\lambda_{\mathrm{TH}} = 5$ dB (see Table II).

2) Influence of ρ_{TH}: Under the condition $d = 1$ m, $N = 4$, SNR $= 30$ dB, two DRRs (0 and 20 dB), and $\lambda_{\mathrm{TH}} = 5$ dB, we examine how the performance of the ICR algorithm varies when ρ_{TH} increases from 0.1 to 0.9, with an interval of 0.1. We use two versions of the ICR algorithm, one with shifted PVL (see cf., Eq. (27)) and one without shifting (see cf. Eq. (23)). We refer to them as ICR-shift and ICR-noshift, respectively. Fig. 7 depicts the F-scores obtained by these two ICR algorithms. In general, both algorithms perform better in low reverberation than in high reverberation. In low reverberation (DRR $= 20$ dB), ICR-shift and ICR-noshift perform similarly for all ρ_{TH}: they achieve almost perfect results when $\rho_{\mathrm{TH}} < 0.5$ and their performance degrades quickly with increasing ρ_{TH} when $\rho_{\mathrm{TH}} > 0.5$. In high reverberation (DRR $= 0$ dB), ICR-shift is less sensitive to the value of ρ_{TH} than ICR-noshift, whose performance improves quickly with ρ_{TH} and peaks at $\rho_{\mathrm{TH}} = 0.6$. The optimal ρ_{TH} value for ICR-noshift depends on the reverberation density greatly. In contrast, with shifting processing, ICR-shift performs more robustly against reverberation and obtain a high F-score at $\rho_{\mathrm{TH}} = 0.4$ for both DRRs. This value is close to our choice $\rho_{\mathrm{TH}} = 0.3$ (see Table II).

3) Performance Comparison: At first, we compare the performance of the three algorithms (ICR, DC, DEMIX) for different DRRs (increasing from -10 to 20 dB, with an interval of 5 dB) and numbers of sources ($N \in [3, 6]$), when $d = 1$ m and SNR $= 30$ dB. Fig. 8 shows the resulting F-scores, which increase with DRR. DEMIX performs the worst. ICR performs much better than the other two algorithms when DRR $\geq$0 dB. ICR achieves almost perfect results for all N when DRR ≥ 5 dB. All algorithms perform poorly in high reverberation with DRR ≤ -5 dB, achieving F-scores smaller than 0.5. We assume that the algorithms fail in this case when their F-scores are

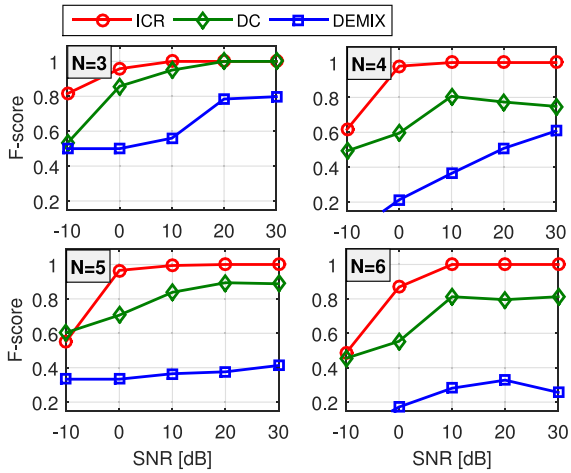


Fig. 9. Performance (F-score) comparison of the source counting algorithms for different numbers of sources N and SNRs. The inter-microphone distance is 1 m, DRR = 20 dB.

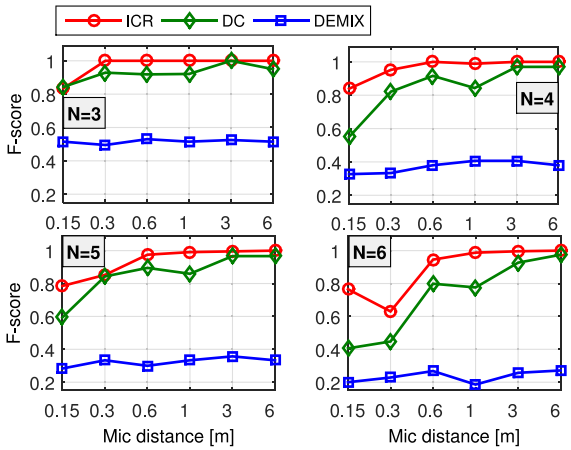


Fig. 10. Performance (F-score) comparison of the source counting algorithms for different numbers of sources N and inter-microphone distances. DRR = 10 dB, SNR = 10 dB.

low enough (e.g., <0.5). In high reverberation, the reverberant part may be stronger than the direct part. As a result, the highest peak of the GCC function may not denote the true TDOA of the source, leading to the failure of ICR. The poor performance of DEMIX is due to its clustering operation solely on the signal amplitude. In our simulation, all the sources have similar amplitude and thus can not be distinguished using this information alone.

Next, we compare the performance of the three algorithms for different SNRs (increasing from -10 to 30 dB, with an interval of 10 dB) and numbers of sources ($N \in [3, 6]$), when $d = 1$ m and DRR = 20 dB. Fig. 9 shows the resulting F-scores: the performance of all algorithms improves with SNR. ICR performs the best and DEMIX performs the worst. When SNR $\geq$ 10 dB, ICR performs almost perfectly for all N.

Moreover, we compare the performance of the three algorithms for different inter-microphone distances ($d \in \{0.15, 0.3, 0.6, 1, 3, 6\}$ m) and numbers of sources ($N \in [3, 6]$), when DRR = 10 dB and SNR = 10 dB. Fig. 10 shows the

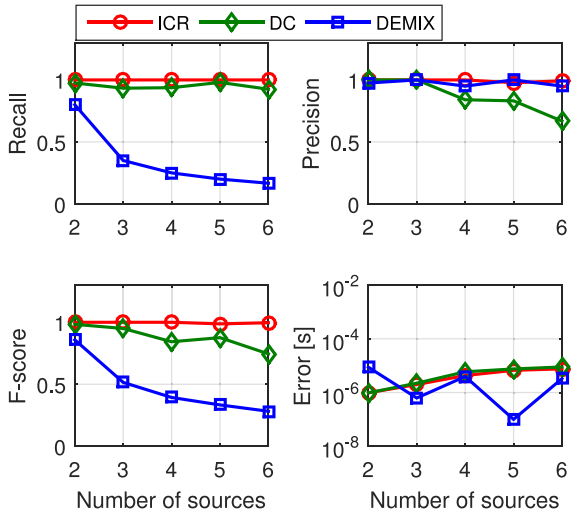


Fig. 11. Performance (recall, precision, F-score and TDOA estimation error) comparison of the source counting algorithms for different numbers of sources. The inter-microphone distance is 1 m, DRR = 10 dB, SNR = 10 dB.

resulting F-scores. DEMIX fails in almost all testing cases, and ICR performs better than DC in most testing cases. The performance of ICR and DC degrades when the inter-microphone distance decreases, which leads to smaller TDOA difference between spatially separated sources. When $d \geq 0.6$ m, ICR achieves almost perfect results for all N.

Finally, we compare the performance in terms of recall rate, precision rate, F-score and localization accuracy of the three algorithms for different numbers of sources ($N \in [2, 6]$), when $d = 1$ m, DRR = 10 dB and SNR = 10 dB (see Fig. 11). ICR and DC achieve a recall rate close to 1 for all testing cases. ICR achieves a precision rate close to 1 for all testing cases, while DC achieves a precision rate which decreases with increasing N. Unlike ICR, DC tends to overestimates the number of sources. Although DEMIX achieves a precision rate close to 1 in all testing cases, its recall rate drops quickly when increasing N. The F-score shows the rank of the global performance as ICR $>$ DC $>$ DEMIX. For localization accuracy, all three algorithms achieve a TDOA estimation error below 10^{-5} s for correctly detected sources. ICR and DC performs the same for localization accuracy, with the TDOA estimation error increasing with N.

E. Results From Image-Source Based Room Impulse Response

In addition to artificial room impulse responses, we also use the image-source method [50] to simulate the room impulse response in an enclosure of size 7 m $\times$ 9 m $\times$ 3 m. The microphones are placed in the center of the enclosure and 1 m apart. The sources are placed $d_{\text{ms}} = 2$ m and 4 m away from the middle of the microphone pair. Seven source directions from 0° to 180°, with an interval of 30°, are considered. All the microphones and sources are placed 1.3 m high. The same speech files and configuration as for the artificial impulse response are used. We consider three scenarios with different reverberation times RT_{60} and microphone-source distance d_{ms}: (a) $\text{RT}_{60} = 100$ ms, $d_{\text{ms}} = 2$ m; (b) $\text{RT}_{60} = 400$ ms, $d_{\text{ms}} = 2$ m; (c) $\text{RT}_{60} = 400$ ms,

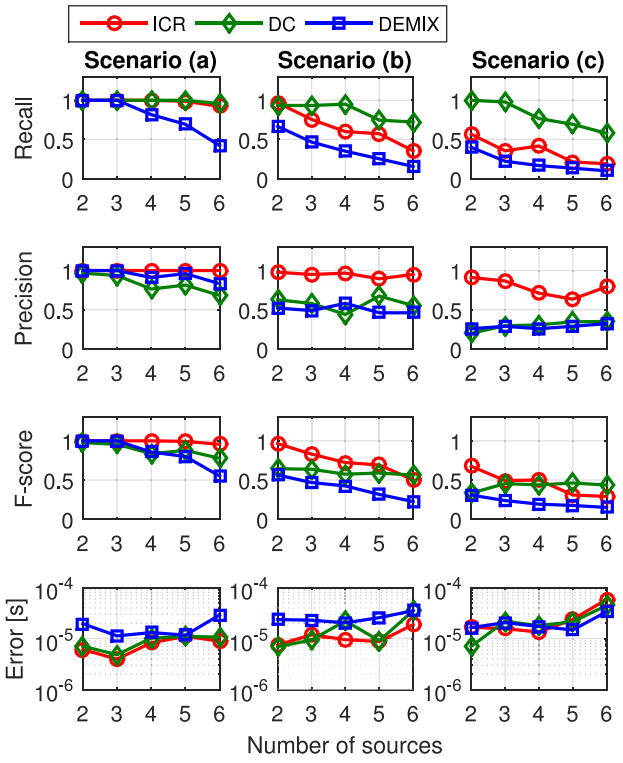


Fig. 12. Performance (recall, precision, F-score and TDOA estimation error) comparison of the source counting algorithms for different numbers of sources simulated with the image-source method. Three scenarios are considered with different reverberation times RT_{60} and microphone-source distance d_{ms}: (a) $\mathrm{RT}_{60} = 100$ ms, $d_{\mathrm{ms}} = 2$ m; (b) $\mathrm{RT}_{60} = 400$ ms, $d_{\mathrm{ms}} = 2$ m; (c) $\mathrm{RT}_{60} = 400$ ms, $d_{\mathrm{ms}} = 4$ m. The inter-microphone distance is 1 m.

$d_{\mathrm{ms}} = 4$ m. The DRRs in the three scenarios are about 6.3, -2.5 and -7.0 dB, respectively.

Fig. 12 shows the four measures obtained by the considered algorithms in different scenarios. For scenario (a), the global performance in terms of F-score can be ranked as ICR $>$ DC $>$ DEMIX. ICR performs well in terms of both recall rate and precision rate; DC performs well in terms of recall rate, but poorly in terms of precision rate; DEMIX performs well in terms of precision rate, but poorly in terms of recall rate. For scenario (b), the global performance in terms of F-score can still be ranked as ICR $>$ DC $>$ DEMIX. The performance of all algorithms degrades as the reverberation time rises to 400 ms. ICR achieves a precision rate close to 1 in all testing cases, but its recall rate decreases evidently when increasing N. The recall rate of DC is close to 1 when $N \leq 4$, and decreases with increasing N when $N > 4$. In comparison to ICR, DC achieves a higher recall rate, but much lower precision rate. For scenario (c), the performance of three algorithms further degrades as d_{ms} is increased to 4 m. ICR degrades more significantly than the other two algorithms, with its recall rate below 0.5 and precision rate below 1 in most cases. Consequently, ICR outperforms DC in terms of F-score when $N \leq 4$, but performs worse than DC when $N \geq 5$. DEMIX still performs the worst. For localization accuracy, the three algorithms obtain similar TDOA errors, around 10^{-5} s, for correctly detected sources in all testing cases.

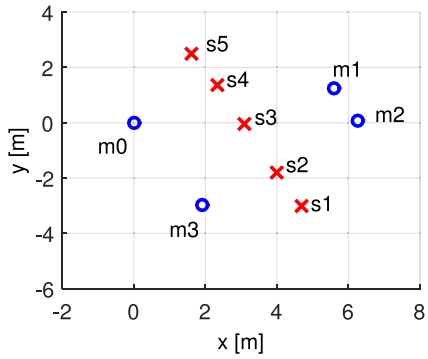


Fig. 13. Geometrical configuration for real recording. The locations of the microphones (m_0–m_3) and sources (s_1–s_5) are denoted by circles and crosses, respectively.

TABLE IV
NUMBER OF SOURCES VERSUS DOA IN THE REAL ENVIRONMENT

Number of Sources	Location
2	s2, s3
3	s1, s3, s5
4	s1, s2, s4, s5
5	s1, s2, s3, s4, s5

F. Results With Data Recorded in a Real Environment

The data are recorded in a quiet public square of about 20 m $\times$ 20 m, with strong reflections from nearby buildings. The positions of four microphones and five sources are shown in Fig. 13. All the sources and microphones are at the same height of 1.5 m. We measure the impulse responses from the sources to the microphones and convolve the impulse responses with speech files to generate the testing data. The DRRs at the microphones are around 5 dB. The same speech files from the simulation are used. We use two microphone pairs, (m1, m2) and (m0, m3), which are about 1.4 m and 3.5 m apart, respectively. For each pair of microphones, five source positions (s1–s5) are considered. The number of sources varies from 2 to 5. The locations of the sources are selected based on the number of sources, as listed in Table IV. For each geometrical configuration we realize 20 instances, where in each instance the speech is randomly selected from the 12 files.

The experimental results are shown in Fig. 14. ICR performs better than DC and DEMIX for both microphone pairs. The performance of ICR and DC degrades quickly as the number of sources increases. There are mainly two reasons for that. First, the linear phase variation is distorted more severely in real environments whose measured acoustic impulse response consists of strong early reflections. Second, the amplitudes of the sources are different, depending on their distances to the microphones. In some cases, the source counting performance may degrade if some sources dominate in the mixtures. In contrast, DEMIX may benefit from the different amplitudes of the sources, by applying clustering to the signal amplitude. For instance, DEMIX achieves a higher F-score for (m_0, m_3) than for

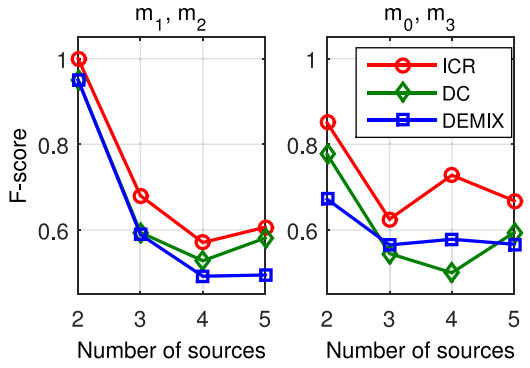


Fig. 14. Performance (F-score) comparison of the source counting algorithms in real environments; (m_1, m_2) are 1.4 m apart while (m_0, m_3) are 3.5 m apart.

(m_1, m_2), where the former pair is farther apart. For (m_0, m_3), DEMIX even outperforms DC in some cases.

G. Computational Complexity

Considering Fig. 1, the computational complexity of the first two blocks (IPD calculation and T-F weighting) of the proposed ICR algorithm remains almost constant in different acoustic environments. The third block ICR involves GCC function calculation, peak detection and contribution removal for each iteration. The computational complexity in each iteration is dominated by the GCC function calculation, which depends on the size of the TDOA search space and the number of valid T-F bins. The number of iterations and the number of valid T-F bins in each iteration depend on the acoustic environment (e.g., the number of sources and reverberation density). The DC algorithm consists of three blocks: IPD calculation, T-F weighting and GCC peak counting. The first two blocks are the same as the ones in the proposed algorithm. The GCC peak counting block only calculates the GCC function once.

We run Matlab code for ICR, DC and DEMIX on an Intel CPU i7 at 3.2 GHz with 16 GB RAM, using the simulated data in Section VI-D. The data length is 10 s with sampling rate 8 kHz. We set SNR $= 30$ dB and DRR $= 20$ dB, and try varying number of sources ($N \in [2, 6]$). Fig. 15(a) depicts the computation time of the considered algorithms, which can be ranked as DEMIX $<$ DC $<$ ICR. The computation time of DEMIX remains almost constant for various N. The computation time of DC decreases with increasing N because, as the number of sources is increased, fewer T-F bins are detected to be one-source active and are taken into account in the GCC function. Fig. 15(b) depicts the computation time of the blocks of the ICR algorithm: IPD + TF-weighting and ICR. The computation time of the IPD + TF-weighting block remains almost constant with N. The computation time of ICR is much higher than IPD + TF-weighting, and does not vary regularly with N.

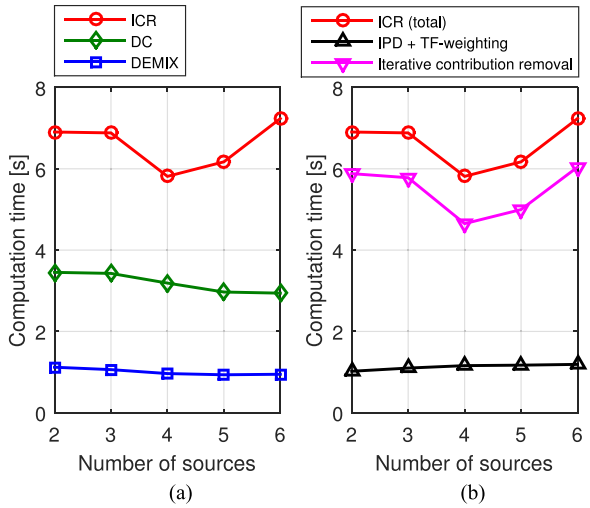


Fig. 15. Computation time of the considered algorithms for 10 s data with varying number of sources. (a) Three algorithms: ICR, DC and DEMIX. (b) Constituent blocks of the ICR algorithm: IPD + TF-weighting and ICR.

VII. Conclusion

We proposed an IPD-based joint source counting and localization scheme for two distant-microphones. The proposed algorithm works in the T-F domain to exploit the nonstationarity and sparsity of audio signals. To count the number of sources from the multiple peaks of the GCC function, we proposed an ICR algorithm that uses GCC and IPD iteratively to detect and remove the T-F bins associated with each source from the IPD plot. Experiments in both simulated and real environments confirmed the effectiveness of the proposed method. Using T-F weighting, the robustness of the proposed algorithm to ambient noise was improved.

The proposed algorithm is suitable for different inter-microphone distances over 0.15 m. In low reverberation, the algorithm can robustly detect up to six sources. In high reverberation (e.g., DRR < 0), the performance degrades significantly, especially when the reverberant part is stronger than the direct part. For the same reason, the performance of the algorithm degrades in real environments with strong early reflections. However, in most cases, the algorithm clearly outperforms other existing approaches.

Since the proposed ICR algorithm only considers the phase information, an interesting future work is to incorporate amplitude information, just as DEMIX does. Instead of hard thresholding, a soft thresholding scheme could also be employed for SNR and coherence weighting. The proposed algorithm only considers static sources with a batch processing and could be extended to moving sources by introducing a frame-by-frame processing scheme and a tracker [54].

References

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," in *Proc. IEEE Symp. Commun. Vehicular Technol. Benelux*, Ghent, Belgium, 2011, pp. 1–6.

[2] M. H. Hennecke and G. A. Fink, "Towards acoustic self-localization of ad hoc smartphone arrays," in *Proc. IEEE Joint Workshop Hands-free Speech Commun. Microphone Arrays*, Edinburgh, U.K., 2011, pp. 127–132.

[3] T. K. Hon, L. Wang, J. D. Reiss, and A. Cavallaro, "Audio fingerprinting for multi-device self-localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1623–1636, Oct. 2015.

[4] L. Wang, T. K. Hon, J. D. Reiss, and A. Cavallaro, "Self-localization of Ad-hoc arrays using time difference of arrivals," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 1018–1033, Feb. 2016.

[5] L. Wang and S. Doclo, "Correlation maximization based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 571–582, Mar. 2016.

[6] M. Brandstein and D. Ward, Eds. *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

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

[8] L. Wang, H. Ding, and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," *EURASIP J. Audio, Speech, Music Process.*, pp. 1–13, 2010, Art. no. 797962.

[9] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using MaxNSR blocking matrix," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1493–1508, Sep. 2015.

[10] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem," in *Independent Component Analysis and Signal Separation*. Berlin, Germany: Springer-Verlag, 2009, pp. 742–750.

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

[12] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 1913–1928, Nov. 2010.

[13] M. Cobos, J. J. Lopez, and D. Martinez, "Two-microphone multi-speaker localization based on a Laplacian mixture model," *Digit. Signal Process.*, vol. 21, no. 1, pp. 66–76, Jan. 2011.

[14] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, vol. 3, pp. 549–557, Mar. 2011.

[15] L. Wang, "Multi-band multi-centroid clustering based permutation alignment for frequency-domain blind speech separation," *Digit. Signal Process.*, vol. 31, pp. 79–92, Aug. 2014.

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

[17] V. V. Reddy, A. W. H. Khong, and B. P. Ng, "Unambiguous speech DOA estimation under spatial aliasing conditions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2133–2145, Dec. 2014.

[18] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[19] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP J. Adv. Signal Process.*, vol. 2003, pp. 1110–1124, 2003.

[20] A. Lombard, Y. Zheng, H. Buchner, and W. Kallermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1490–1503, Aug. 2011.

[21] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[22] F. Nesta and O. Maurizio, "Generalized state coherence transform for multidimensional TDOA estimation of multiple sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 246–260, Jan. 2012.

[23] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Mar./Apr. 2008, pp. 2565–2568.

[24] A. Clifford and J. Reiss, "Calculating time delays of multiple active sources in live sound," in *Proc. 129th Audio Eng. Soc. Convention*, San Francisco, CA, USA, 2010, pp. 1–8.

[25] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, Apr. 1997, vol. 1, pp. 375–378.

[26] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein, D. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, pp. 157–180.

[27] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, Jan. 2011.

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

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

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

[31] E. D. Di Claudio, R. Parisi, and G. Orlandi, "Multi-source localization in reverberant environments by ROOT-MUSIC and clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, Turkey, 2000, vol. 2, pp. 921–924.

[32] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011, pp. 117–120.

[33] M. Kepesi, L. Ottowitz, and T. Habib, "Joint position-pitch estimation for multiple speaker scenarios," in *Proc. Hands-Free Speech Commun. Microphone Arrays*, Trento, Italy, May 2008, pp. 85–88.

[34] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 923–933, May 2013.

[35] S. Gerlach, J. Bitzer, S. Goetze, and S. Doclo, "Joint estimation of pitch and direction of arrival: Improving robustness and accuracy for multi-speaker scenarios," *EURASIP J. Audio, Speech, Music Process.*, vol. 2014, pp. 1–17, 2014.

[36] J. R. Jensen, M. G. Christensen, J. Benesty, and S. H. Jensen, "Joint spatio-temporal filtering methods for DOA and fundamental frequency estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 174–185, Jan. 2015.

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

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

[39] 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," *J. Acoust. Soc. Amer.*, vol. 135, no. 2, pp. 742–753, Feb. 2014.

[40] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 791–803, Nov. 2003.

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

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

[43] C. Kim, C. Khawand, and R. M. Stern, "Two-microphone source separation algorithm based on statistical modeling of angle distributions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 4629–4632.

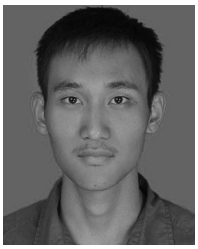
[44] Y. Oualil, F. Faubel, and D. Klakow, "A probabilistic framework for multiple speaker localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 3962–3966.

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

[46] J. Hollick, I. Jafari, R. Togneri, and S. Nordholm, "Source number estimation in reverberant conditions via full-band weighted, adaptive fuzzy c-means clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 7450–7454.

[47] L. Drude, A. Chinaev, T. H. Tran Vu, and R. Haeb-Umbach, "Towards online source counting in speech mixtures applying a variational EM for complex Watson mixture models," in *Proc. 14th Int. Workshop Acoust. Signal Enhancement*, Juan les Pins, France, Sep. 2014, pp. 213–217.

[48] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.
[49] Z. Lu and A. M. Zoubir, "Flexible detection criterion for source enumeration in array processing," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1303–1314, Mar. 2013.
[50] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[51] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.
[52] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[53] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, NY, USA: Wiley, 2004.
[54] O. Cappe, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proc. IEEE*, vol. 95, no. 5, pp. 899–924, May 2007.



Lin Wang received the B.S. degree in electronic engineering from Tianjin University, Tianjin, China, in 2003, and the Ph.D. degree in signal processing from the Dalian University of Technology, Dalian, China, in 2010. From 2011 to 2013, he was an Alexander von Humboldt Fellow at the University of Oldenburg, Oldenburg, Germany. Since 2014, he has been a Postdoctoral Researcher in the Centre for Intelligent Sensing at Queen Mary University of London, London, U.K. His research interests include video and audio compression, microphone array, blind source separation, and 3D audio processing (https://sites.google.com/site/linwangsig).

Tsz-Kin Hon received the B.Eng. degree in electronic and computer engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2006, and the Ph.D. degree in digital signal processing from Kings College London, London, U.K., in 2013. He was a Research Engineer in the R&D of the Giant Electronic Ltd., between 2006 and 2009. He is currently a Postdoctoral Research Assistant in the Centre for Intelligent Sensing at Queen Mary University of London, London, U.K. His research interests include audio and video signal processing, device localization and synchronization, multi-source signal processing, joint time frequency analysis and filtering, acoustic echo cancellation, speech enhancement, and biomedical signal processing.

Joshua D. Reiss received the Bachelor's degrees in both physics and mathematics, and the Ph.D. degree in physics from the Georgia Institute of Technology, Atlanta, GA, USA. He is currently a Reader in Audio Engineering with the Centre for Digital Music in the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K. He is a Member of the Board of Governors of the Audio Engineering Society, and cofounder of the company MixGenius, now known as LandR. He has published more than 100 scientific papers and serves on several steering and technical committees. He has investigated sound synthesis, time scaling and pitch shifting, source separation, polyphonic music transcription, loudspeaker design, automatic mixing for live sound, and digital audio effects. His primary focus of research, which ties together many of the above topics, is on the use of state-of-the-art signal processing techniques for professional sound engineering.

Andrea Cavallaro received the Ph.D. degree in electrical engineering from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2002. He was a Research Fellow with British Telecommunications in 2004. He is currently a Professor of multimedia signal processing and the Director of the Centre for Intelligent Sensing at the Queen Mary University of London, London, U.K. He has authored more than 150 journal and conference papers, one monograph on Video Tracking (Hoboken, NJ, USA: Wiley, 2011), and three edited books, Multi-Camera Networks (Amsterdam, The Netherlands: Elsevier, 2009), Analysis, Retrieval and Delivery of Multimedia Content (New York, NY, USA: Springer, 2012), and Intelligent Multimedia Surveillance (New York, NY, USA: Springer, 2013). He is an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING and a Member of the editorial board of the *IEEE MultiMedia Magazine*. He is an Elected Member of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee, and is the Chair of its Awards Committee, and an Elected Member of the IEEE Circuits and Systems Society Visual Communications and Signal Processing Technical Committee. He served as an Elected Member of the IEEE Signal Processing Society Multimedia Signal Processing Technical Committee, as an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA AND IEEE TRANSACTIONS ON SIGNAL PROCESSING, as an Associate Editor and as an Area Editor of *IEEE Signal Processing Magazine*, and as a Guest Editor of eleven special issues of international journals. He was the General Chair for the IEEE/ACM ICDSC 2009, BMVC 2009, M2SFA2 2008, SSPE 2007, and IEEE AVSS 2007. He was Technical Program Chair of the IEEE AVSS 2011, the European Signal Processing Conference in 2008, and WIAMIS 2010. He received the Royal Academy of Engineering Teaching Prize in 2007, three Student Paper Awards on target tracking and perceptually sensitive coding at the IEEE ICASSP in 2005, 2007, and 2009, respectively, and the Best Paper Award at IEEE AVSS 2009.