Indoor Multi-Speaker Localization based on Bayesian Nonparametrics in the Circular Harmonic Domain

Kunkun SongGong, *Student Member, IEEE*, Huawei Chen, *Member, IEEE*, and Wenwu Wang, *Senior Member, IEEE*

***Abstract*—Circular microphone arrays have been used for multi-speaker localization in computational auditory scene analysis, for their high flexibility in sound field analysis, including the generation of frequency-invariant eigenbeams for wideband acoustic sources. However, the localization performance of existing circular harmonic approaches, such as circular harmonics beamformer (CHB) depends strongly on the physical characteristics (such as shape) of sensor arrays, and the level of uncertainties presented in acoustic environments (such as background noise, room reverberation, and the number of sources). These uncertainties may limit the performance or practical application of the speaker localization algorithms. To address these issues, in this paper, we present a new indoor multi-speaker localization method in the circular harmonic domain based on the acoustic holography beamforming (AHB) technique and the Bayesian nonparametrics (BNP) method. More specifically, we use the AHB technique, which combines the delay-and-sum beamforming with acoustic-holography-based virtual sensing, to generate direction of arrival (DOA) measurements in the time-frequency (TF) domain, and then design a BNP algorithm based on the infinite Gaussian mixture model (IGMM) to estimate the DOAs of the individual sources without the prior knowledge about the number of sources. These estimates may degrade in the presence of room reverberation and background noise. To address this issue, we develop a robust TF bin selection and permutation method on the basis of mixture weights, using power, power ratio and local variance estimated at each TF bin. Experiments performed on both simulated and real-data show that our method gives significantly better performance, than four recent baseline methods, in a variety of noise and reverberation levels, in terms of the root-mean-square error (RMSE) of the DOA estimation and the source detecting success rate.**

***Index Terms*—Multi-speaker localization, Bayesian nonparametrics (BNP), circular harmonics, direction of arrival (DOA) estimation, microphone array signal processing.**

I. Introduction

SPEAKER localization, or the direction of arrival (DOA) estimation of speech sources, using acoustic sensor arrays, has received extensive attention in recent years. As an important and active research topic in audio signal processing, it has a wide variety of practical applications, e.g., in video teleconferencing, where the knowledge of the location of a speaker helps steer a camera to focus on the speaker [1]; in robotics, where microphones can be installed on robots to estimate human positions for effective human-robot interaction [2]; in smart home design, where the locations of the sources can be taken into account to facilitate the design of smart devices [3]; and in hearing aids, speech source location can be used to enhance its intelligibility in a noisy environment [4], among many others [5]–[7].

Existing approaches for DOA estimation can be classified approximately into four categories, namely, the steered-response power (SRP) beamforming methods [8], [9], the subspace-based methods [10]–[12], the time-difference-of-arrival (TDOA) estimation methods and the intensity-based methods. The SRP beamforming methods [8] are an intuitive solution for DOA estimation, by steering the beam over DOAs and analyzing the highest power. Owing to their limitations on spatial resolution, SRP methods may fail to localize the speakers that are close to each other [2]. To overcome this limitation, a method based on the spatial sound presence probability (SSPP) was proposed in [9], where spatial probabilities based on a relative transfer function (RTF) correlation feature are used to incorporate knowledge of the anechoic RTFs in the directions of interest for the localization of multiple simultaneously active sources in adverse environments. The subspace-based methods, including the popular methods e.g. multiple signal classification (MUSIC) [10] and estimation of signal parameters via rotational invariance techniques (ESPRIT) [11], utilize the eigen-decomposition of the covariance matrix of the received signals at microphones to achieve good DOA resolutions. They are mainly applicable to source localization in anechoic environments, whereas the coherent signal subspace (CSS) based framework presented in [12] can be used to deal with reverberant scenarios. The TDOA based methods are often exploiting the Generalized Cross-Correlation PHAse Transform (GCC-PHAT) [13]. However, they are designed originally for single source localization and the localization results may be sensitive to room reverberation due to the free-field plane-wave model it assumes [14]. The intensity-based methods determine the magnitude and direction of the transport of acoustic energy, related to the DOA of a sound wave [15]. Unfortunately, in practice, it is difficult to measure particle velocity, although attempts have been made to use the finite difference method with two microphone arrays [16].

In common with most speaker localization methods, how-

Manuscript received August 1, 2020; revised January 19, 2021 and April 23, 2021; accepted May 6, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61971219 and 61471190 and in part by the State Key Laboratory of Acoustics, Chinese Academy of Sciences, under Grant No. SKLA202015. (*Corresponding author: Huawei Chen*)

K. SongGong and H. Chen are with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (email: sgkk@nuaa.edu.cn, hwchen@nuaa.edu.cn).

W. Wang is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: w.wang@surrey.ac.uk)

ever, there are still several challenges remaining: (1) the characteristics of the array, particularly the small-sized array, may affect the localization performance or limit its practical application, (2) localizing multiple and simultaneously active speakers is an extremely important and challenging issue, especially when the number of sources is unknown, and (3) the presence of room reverberation and background noise in the microphone measurements further complicates the problem and degrades the accuracy of DOA estimation.

In the last several years, modal signal processing using sensor array [17] has received increasing attention. The reason for this is that it can provide a frequency-invatiant eigenbeam and be used for localizing wideband source without a narrowband assumption underlying the traditional signal model [18]. The authors of [19] developed the circular harmonics beamformer (CHB), which belongs to a more recent class of methods often referred to as eigenbeamforming. This method is shown to achieve better resolution and sidelobe properties than delay-and-sum beamforming by selectively processing a different number of phase modes. In particular, a time-frequency (TF) - CHB [20] was reported to have outperformed the eigenbeam (EB) - ESPRIT [21] in localizing multiple sources under a high level of reverberation and noise. Despite being straightforward, the localization performance of these methods depends strongly on the physical characteristics of the sensor array. In [15], a method using the pseudointensity vector (PIV) is designed for the localization of a single source, which uses sound field information with low spatial resolution. As an extension of PIV, the Subspace PIV (SS-PIV) [22] uses the low order spatial information of signal subspace for the localization of multiple sources. However, this method is sensitive to noise and room reverberation. In [23], MUSIC with direct-path dominance (DPD) test is used to improve multi-source localization in highly reverberant environments by exploiting the sparsity of speech in the TF domain. The DPD test aims to identify TF bins that contain significant contributions from a single source, i.e. the direct signal. However, as described in [23], the DPD test is reported to be satisfied in only 3% of the TF regions. As a result, there may not be any DOA estimates in some time frames, which is problematic in practice.

Although the multi-speaker localization methods based on circular harmonics are promising as aforementioned, they still suffer from several limitations in practical applications: (1) the accuracy of DOA estimation may be limited by the shape of sensor arrays, for instance the number of transducers, the radius and whether the sensors are mounted on a scatterer such as a rigid cylinder or a sphere. Specifically, the TF-CHB method is improved by increasing the number of sensors and the radius of the array simultaneously as this increases the maximum order that can be used [20]. However, the requirements said above are not all met in practical applications, because the number of sensors and the array radius cannot be increased without limit; (2) all the aforementioned multi-source localization approaches have a common issue that the maximum number of sources has to be specified in advance, which is often unknown in practical situations. Since *a priori* knowledge about the acoustic environment including the source number is often difficult to obtain, an all-round method applicable to a wide range of environments is desirable [24]; and (3) although the existing circular harmonic DOA estimation methods perform well in low or moderate noise conditions [19], they are known to be sensitive to noise and reverberation [14], [20], and hence degrade in these adverse conditions for multi-speaker scenarios.

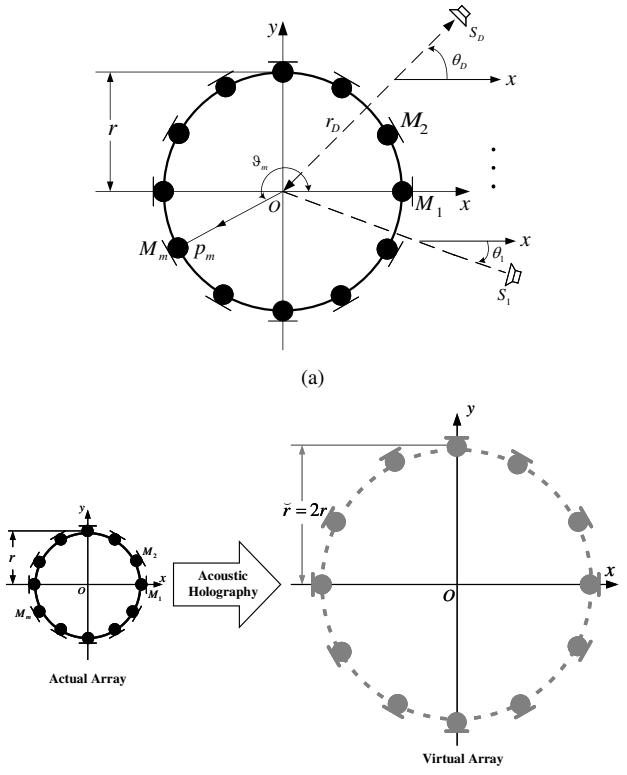
In this paper, we present an indoor multi-speaker localization method in the circular harmonic domain. We firstly develop an acoustic holography beamforming (AHB) technique in the TF domain, which is in some sense analogous to the technique commonly used for capturing the evanescent waves to enhance spatial resolution [25], [26]. The acoustic holography method was introduced originally in [27] and conceived for array measurements in the near-field of a source to predict the sound field close to it, with the aim to visualize the source radiation characteristics. Recently, the combination of acoustic holography and beamforming has been examined by Fu *et al.* [28] for visualization of sound sources with high temperature. Different from these works, here we develop an AHB technique in the TF domain, where delay-and-sum beamforming is combined with acoustic-holography-based virtual sensing, for modelling multi-speaker localization, thereby overcoming the limitations of the conventional circular harmonic DOA estimation methods, which are strongly influenced by practical constraints of array characteristics as aforementioned. To our knowledge, this is the first time that AHB is used for multi-speaker localization. Secondly, we develop a Bayesian non-parametrics (BNP) algorithm [24] for the DOA estimation of the sources under the AHB model. The BNP algorithm is able to determine the complexity of the model without the prior knowledge about the number of sources. More specifically, the BNP clustering method is used to address this problem by assuming that there is an infinite number of latent clusters, but only a finite number of them is used to generate the observed data [29]. Furthermore, the BNP models allow the complexity to grow as more data are observed, such as in [30]. Thus, we can overcome the source number uncertainty issue using a BNP method. In addition, we modify the BNP formulation to allow more emphasis on the DOA observations by using the mixture weights, which can eliminate the permutation problem introduced by noise and reverberation. The results of simulations and real-data experiments in noisy and reverberant environments show the superior performance of our proposed multi-speaker localization method when compared with the existing circular harmonic methods.

The remainder of this paper is structured as follows. Section II provides a problem formulation. Section III presents an overview of the proposed method, which contains three main stages as detailed in Sections IV, V and VI, respectively. Section IV presents the exploitation of the TF-AHB technique in the circular harmonic domain for generating the TF measurements. Section V presents a BNP model and an inference algorithm for the estimation of the number of sources and their DOAs. Section VI presents a technique to improve the parameter estimation of the BNP model in reverberant and noisy conditions based on the use of mixture weight and TF bin selection. The simulation and real-data experimental results are discussed in Section VII. Finally, the conclusions are drawn in Section VIII.

II. Problem Formulation

A uniform circular array (UCA) consisting of M omnidirectional sensors is adopted for sound source capture as shown in Fig. 1(a). This is mainly due to its simple and compact structure, and also its DOA estimation range from $-180°$ to $180°$. The geometric center of the UCA is chosen as the origin of the coordinate system, the radius of the array is r, and the azimuth angle of each sensor is ϑ_m, namely

$$\vartheta_m \quad = \quad (m-1)\frac{2\pi}{M}, \qquad (1 \leq m \leq M). \tag{1}$$



(a)

(b)

Fig. 1. (a) Configuration of the uniform circular sensor array. (b) The virtual array used in acoustic holography, where the gray points denote the positions of the virtual array elements. Note that the figures are not drawn to scale.

Suppose that D far-field speech sources in a reverberant enclosure impinge on the array. Herein, the DOAs are defined with respect to the positive x-axis, which implies $\theta_d \in [-180°, 180°), d = 1, \ldots, D$. The source signal received at the mth sensor can be modeled as

$$p_m(\tau) \quad = \quad \sum_{d=1}^{D} h_{dm}(\tau) * s_d(\tau) + v_m(\tau), \tag{2}$$

where $s_d(\tau)$ is the signal induced by one of D speech sources at r_d distance from the centre O of microphone array, $h_{dm}(\tau)$ is the room impulse responses (RIRs) from the dth source to the mth sensor, τ is the discrete time index, $*$ is the convolution operator, and $v_m(\tau)$ is the additive background noise.

In the short-time Fourier transform (STFT) domain, (2) can be transformed to

$$P_m(k,t) \quad = \quad \sum_{d=1}^{D} H_{dm}(k,t)S_d(k,t) + V_m(k,t), \tag{3}$$

where $k = 2\pi f/c$ is the wavenumber, t is the time frame index, f is the frequency, c is the speed of sound, and $P_m(k,t)$, $S_d(k,t)$, $H_{dm}(k,t)$, and $V_m(k,t)$ are the STFT of $p_m(\tau)$, $s_d(\tau)$, $h_{dm}(\tau)$, and $v_m(\tau)$, respectively.

Speech signals, in general, are considered sparse in the TF domain [14], and as a result, the speech sources from multiple speakers will not be substantially overlapping in the TF domain. In other words, at each TF bin, it could be assumed that only one source is dominant, i.e., the probability of one source presenting at this TF bin is higher than those of other sources. With the sparsity assumption, (3) can be further simplified as follows

$$P_m(k,t) \quad \approx \quad H_{dm}(k,t)S_d(k,t) + V_m(k,t). \tag{4}$$

Such an assumption has been exploited in source localization [31] and source separation [6], [32], [33]. In a reverberant environment, the presence of room reverberation increases the chance of source overlap in the TF domain. Nevertheless, we characterize the presence of each source with a probability showing how likely it occurs at each TF bin. Due to such a representation, our method allows the presence of reverberation and overlap between sources in the microphone signals to be described using a probabilistic model and quantified with source presence probabilities at each TF bin. We argue that the sparsity assumption is less restrictive than the W-disjoint orthogonality (W-DO) assumption [34], which is less likely to hold for multiple sources and reverberant environments.

Our objective is to estimate the DOAs θ_d of the speech sources based on the received mixture signals $P_m(k,t)$ in an indoor environment. To this end, we propose a new method where the AHB technique is used to generate the DOA measurements for all the TF bins, and then the BNP method is developed to estimate the DOAs of all the sources, without the prior knowledge about the number of sources. We also present a new and robust method for estimating the parameters of the BNP model based on mixture weights, in the presence of noise and room reverberation.

III. Overview of the Proposed Method

The proposed method, as shown in Fig. 2, is composed of three core stages. First, given the microphone signals $p_m(\tau)$, the DOA measurements in the TF domain, i.e. $\hat{\Theta} = \{\hat{\theta}(k,t)\} = \{\hat{\theta}_1, \cdots, \hat{\theta}_i, \cdots, \hat{\theta}_I\}$, where I is the total number of TF bins, are generated by the AHB approach with the steps shown in the blue boxes in Fig. 2, which include the conversion of $p_m(\tau)$ to a TF domain representation $P_m(k,t)$, and the formation of acoustic holography beamforming in order to obtain the measurements $\hat{\Theta}$.

Second, an Infinite Gaussian Mixture Model (IGMM) of l components with hyperparameters $\mathbf{\Psi}_l$ is used to model $\hat{\Theta}$. The BNP method is used to estimate, in an iterative and alternating manner, the model parameters $\mathbf{\Psi}_l$, the posterior probability of k_i (i.e. the shift required to unwrap the original $\hat{\theta}_i$ to a range outside $[-\pi, \pi)$, $i = 1, ..., I$), and the posterior probability of class label z_i (i.e. indicating to which mixture component l that $\hat{\theta}_i$ belongs). Here, Gibbs sampling is used to simplify the computation of the posterior probability of z_i. This stage corresponds to the black boxes in Fig. 2.

Third, a refined set of DOA measurements $\hat{\Theta}'$ is obtained by selecting Q most reliable measurements from $\hat{\Theta}$, i.e. those with highest mixture weights $W(k,t)$ computed at corresponding TF bins. The measurements in $\hat{\Theta}'$ are further re-aligned to mitigate permutation ambiguities, leading to $\hat{\Theta}''$, which are then used as inputs to the BNP algorithm to refine the estimate of the model parameters $\mathbf{\Psi}_l$ and the posterior probabilities z_q and k_q, $q=1,...,Q$. This stage corresponds to the green boxes in Fig. 2. Finally, the number of mixture components l, i.e. estimated D and their DOAs, i.e. the estimated $\hat{\theta}_d$, $d=1,...,Q$, can be obtained upon the convergence of BNP algorithm, after all the $\hat{\theta}(k,t)$ have been clustered into one of the l mixture components, with l being updated in each iteration.

The details of the proposed method are presented in the next three sections.

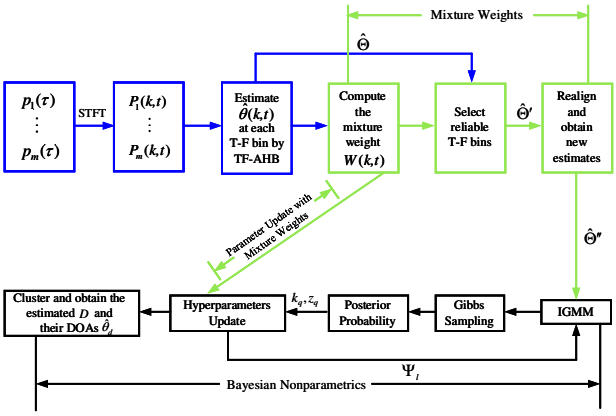


Fig. 2. Block diagram showing the main stages of processing in the proposed method.

IV. Time-Frequency Acoustic Holography Beamforming

Acoustic holography is a sound visualization technique that makes it possible to reconstruct the entire acoustic field via expanding the measured sound field into a series of basis functions [26]. In this paper we focus on a circular array, making use of the fact that the sound field can be predicted on different radii by using Bessel functions that account for the propagation in the radial direction.

We consider the scenario where the plane waves travelling from the speakers in a two-dimensional (2D) plane (e.g. xy-plane) are captured by a UCA at the same plane. For the case of 3D plane, a spherical array [35] is often used, but is not considered here. The sound field in the 2D plane can be decomposed into a Fourier series in terms of the azimuth coordinate, ϑ. After applying the boundary conditions (i.e. the sound field at the origin must be finite) [36], the sound pressure at any point of the array (e.g. at an arbitrary radius R) can be written as

$$P(kR,\vartheta) \quad = \quad \sum_{n=-\infty}^{\infty} A_n J_n(kR) e^{jn\vartheta}, \tag{5}$$

where $j = \sqrt{-1}$, n is the number of harmonics, $J_n(\cdot)$ is the nth-order Bessel function of the first kind, and A_n is an expansion coefficient of the nth term. With (5), the sound pressure at an arbitrary point of the sound field can be determined via acoustic holography [36], using different values of A_n. Thus, the pressure at the UCA, namely, at the array radius $R = r$ (see Fig. 1(a)), can be computed as

$$P(kr,\vartheta) \quad = \quad \sum_{n=-\infty}^{\infty} A_n J_n(kr) e^{jn\vartheta}, \tag{6}$$

where the terms $e^{jn\vartheta}$, often referred to as the circular harmonics, form a set of orthogonal functions,

$$\frac{1}{2\pi}\int_0^{2\pi} e^{jn\vartheta}(e^{ju\vartheta})^* \, d\vartheta \quad = \quad \delta_{nu}, \tag{7}$$

where δ_{nu} is the Kronecker delta function, which equals unity when $n = u$ and zero otherwise, and $(\cdot)^*$ denotes the complex conjugate.

The coefficients A_n can be computed by making use of the continuous orthogonality property of the circular harmonics given in (7) and retrieved by multiplying each side of (6) by a complex conjugated circular harmonic and then integrated over the entire circle, from 0 to 2π, as follows

$$A_n \quad = \quad \frac{\frac{1}{2\pi}\int_0^{2\pi} P(kr,\vartheta) e^{-jn\vartheta} \, d\vartheta}{J_n(kr)}. \tag{8}$$

This expression represents a continuous integral of the sound pressure. However, with an array of M microphones, the sound pressure is sampled at discrete positions, rather than in a continuous circle. This implies that the coefficients defined in (8) need to be approximated by a finite summation

$$\int_0^{2\pi} P(kr,\vartheta) e^{-jn\vartheta} \, d\vartheta \approx \sum_{m=1}^{M} \frac{2\pi}{M} P(kr,\vartheta_m) e^{-jn\vartheta_m}, \tag{9}$$

Therefore, the coefficients A_n can be calculated as

$$\tilde{A}_n \quad = \quad \frac{\frac{1}{M}\sum_{m=1}^{M} P(kr,\vartheta_m) e^{-jn\vartheta_m}}{J_n(kr)}. \tag{10}$$

In theory, the sound pressure is represented by an infinite number of Fourier coefficients. In practice, however, the number of coefficients used can be truncated to a finite value that follows $N = \lceil kr \rceil + 1$, where $\lceil \cdot \rceil$ is a ceiling function, as the contributions from the terms associated with the orders higher than n are very small [42]. According to the theory for sampling in space [17], the number of sensors required should satisfy $M > 2N$, in order to capture the sound field up to an order N. Thus, when using M microphones, a wavefield can be decomposed into a combination of harmonic components with a limited number of orders, namely, $N = \begin{cases} M/2-1, & M \text{ even} \\ (M-1)/2, & M \text{ odd} \end{cases}$.

As a result, the sound pressure described in (5) can be truncated as

$$\tilde{P}(kR,\vartheta) \quad = \quad \sum_{n=-N}^{N} \tilde{A}_n J_n(kR) e^{jn\vartheta}, \tag{11}$$

where $\tilde{A}_n$ are the truncated coefficients denoted in (10), which is calculated by the radius of actual array r.

The speech sources can be localized using a beamforming technique. With a delay-and-sum beamformer (DSB), the output at the UCA [19] can be expressed as

$$B(kR,\theta) = \frac{1}{M}\sum_{g=1}^{M} P(kR,\vartheta_g,\theta_d)e^{jkR\cos(\vartheta_g-\theta)} = \frac{1}{M}\sum_{g=1}^{M} P(kR,\vartheta_g)e^{jkR\cos(\vartheta_g-\theta)}, \quad (12)$$

where $P(kR,\vartheta_g,\theta_d)$ is the pressure measured at the gth microphone due to a plane wave with origin at θ_d, for simplicity, $P(kR,\vartheta_g,\theta_d) = P(kR,\vartheta_g)$, and $\theta \in [-\pi,\pi)$.

Next, we combine acoustic holography and beamforming to estimate the DOA of sources in the circular harmonic model. As shown in Fig. 1(a), the pressure is measured with a UCA of radius r and M microphones placed at ϑ_m. Using acoustic holography, the pressure at a virtual array with a larger radius $\breve{r}$, as presented in Fig. 1(b), can be predicted. In the present study, the number of virtual microphones and their azimuth angles are the same as those for the actual array, namely $\vartheta_m = \vartheta_g$. In fact, as shown in [36], we know that the position of the microphones is not that relevant as long as the distance between microphones remains constant. This makes sense since UCAs are practically shift-invariant, i.e., the beamforming pattern is the same regardless of the focusing direction [26].

The pressure predicted with acoustic holography by evaluating (11) at $(R=\breve{r},\vartheta=\vartheta_g)$, i.e. $\tilde{P}(k\breve{r},\vartheta_g)$, is then used as the input to the beamformer. The coefficients $\tilde{A}_n$ given in (10) are obtained by the pressure measured with the microphones of the actual array at (r,ϑ_m). Therefore, the acoustic holography beamforming (AHB) with the radius $\breve{r}$ can be obtained by substituting (10) and (11) into (12),

$$B_{\mathrm{AH}}(k\breve{r},\theta) = \frac{1}{M^2}\sum_{g=1}^{M}\sum_{n=-N}^{N}\sum_{m=1}^{M} P(kr,\vartheta_m)\times \frac{J_n(k\breve{r})}{J_n(kr)}e^{j(n(\vartheta_g-\vartheta_m)+k\breve{r}\cos(\vartheta_g-\theta))}, \quad (13)$$

where $N = \lceil k\breve{r}\rceil + 1$, up to a maximum value $M/2-1$. In our work, we set $M=12$.

Using a property of the Bessel functions, i.e. $e^{jk\breve{r}\cos(\varpi)} = \sum_{n=-N}^{N} j^n J_n(k\breve{r})e^{jn\varpi}$, we can establish the relationship between the AHB and conventional CHB as follows

$$B_{\mathrm{AH}}(k\breve{r},\theta) = \frac{1}{M}B_{\mathrm{CH}}(kr)[J_n(k\breve{r})]^2, \quad (14)$$

where B_{CH} is the CHB output [19], expressed as

$$B_{\mathrm{CH}}(kr)=\frac{1}{M}\sum_{m=1}^{M}P(kr,\vartheta_m)\sum_{n=-N}^{N}\frac{1}{(-j)^n J_n(kr)}e^{-jn(\vartheta_m-\theta)}. \quad (15)$$

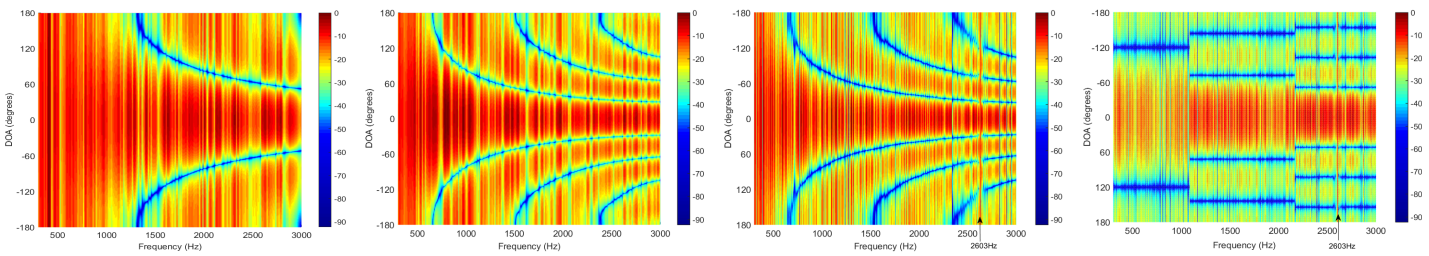
Here, we further clarify the difference between the AHB and conventional CHB. The conventional CHB is based on the actual array with a fixed radius, while the AHB exploits the technique of acoustic holography to predict the sound pressure at different radii of the sound field that would be captured by a virtual array. As a result, it allows the array to work at different radii with the same number of microphones, positioned similarly to those in the actual array, as in Fig. 1(b).

An advantage of AHB over the CHB and DSB is depicted in Fig. 3, which shows the beampatterns provided by a DSB method with the actual array ($r=0.05$ m and $r=0.1$ m), an AHB method with the virtual array ($\breve{r}=2r=0.1$ m), and a CHB method [19] with the actual array ($r=0.05$ m), using a UCA with $M=12$ for DOA $\theta_d=0°$. Comparing Fig. 3 (c) with Fig. 3 (a) and (d), we can notice that the beampattern provided by the virtual array for the radius $r=0.05$ m is more directive, as expected from the theory. From Fig. 3 (b) and (c), we can observe that the virtual array is almost the same as the actual array with radius $r=0.1$ m.

We can also notice a singularity (zero) at the frequency 2603 Hz in Fig. 3 (c) and (d), due to the use of the Bessel function in the denominator [19]. This is the well-known forbidden frequency problem or singularity problem, as shown in [37]. This problem can be avoided by using a cylindrical array or directional microphones, however, in practice, this may not be realistic and flexible to achieve [38]. In our work, we avoid this problem by limiting the upper frequency at 2500 Hz, which also mitigates the spatial aliasing problem (occurring at 3300 Hz for $\breve{r}=0.1$ m). Here, we choose $\breve{r}=2r=0.1$ m for two reasons. Firstly, as shown in Fig. 3, it provides a better beampattern than $r=0.05$ m and almost the same pattern as $r=0.1$ m. Secondly, as in [39], the actual array with a radius 0.1 m is usually chosen for convenience.

As aforementioned, there is one dominant sound source at each TF. Note that, $P(kr,\vartheta_m) = P_m(k,t)$. Therefore, the equation (13) can be written as

$$B_{\mathrm{AH}}(k\breve{r},t,\theta) = \frac{1}{M^2}\sum_{g=1}^{M}\sum_{n=-N}^{N}\sum_{m=1}^{M} P_m(k,t)\times \frac{J_n(k\breve{r})}{J_n(kr)}e^{j(n(\vartheta_g-\vartheta_m)+k\breve{r}\cos(\vartheta_g-\theta))}, \quad (16)$$



(a) (b) (c) (d)

Fig. 3. Beampatterns for DOA $\theta_d=0°$ using a UCA with $M=12$. (a) The actual array with $r=0.05$ m by DSB. (b) The actual array with $r=0.1$ m DSB. (c) The virtual array with $\breve{r}=2r=0.1$ m by AHB. (d) The actual array with $r=0.05$ m by CHB.

For simplicity, we denote $B_{\rm AH}(k\check{r},t,\theta)=B_{\rm AH}(k,t,\theta)$.

The TF-AHB output will have maximum power at its corresponding arrival direction. Thus, the DOA estimate $\hat{\theta}(k,t)$ at each TF bin can be represented as

$$\hat{\theta}(k,t) \;=\; \arg\,\max_{\theta}|B_{\rm AH}(k,t,\theta)|^2. \qquad (17)$$

V. BNP for DOA Estimation of Unknown Number of Sources

In practice, we often need to localize multiple speakers without the prior knowledge about the number of speakers, i.e., D. To address this problem, we develop a BNP method by modelling $\hat{\Theta}=\{\hat{\theta}(k,t)\}=\{\hat{\theta}_1,\cdots,\hat{\theta}_i,\cdots,\hat{\theta}_I\}$ with an IGMM, and estimate D and θ_d $(d=1,...,D)$ via an iterative inference process alternating between the update of the model parameters and the estimation of posterior probability of the class labels z_i of each $\hat{\theta}_i$.

A. Probabilistic Modelling of DOA Measurements

We model $\hat{\Theta}$ with an IGMM [30] containing l components with l being unknown. The Probability Density Function (PDF) of $\hat{\theta}_i$ generated by the mixture component l is given by

$$p(\hat{\theta}_i|\mu_l,\sigma_l^2,k_i) \;=\; \frac{1}{\sqrt{2\pi\sigma_l^2}}e^{\frac{-(\hat{\theta}_i+2\pi k_i-\mu_l)^2}{2\sigma_l^2}}, \qquad (18)$$

where μ_l and σ_l^2 are the mean and variance of the Gaussian component l, and k_i is an integer parameter accounting for the shift of $\hat{\theta}_i$. The original DOAs measurements $\hat{\theta}_i$ are wrapped within the range of $[-\pi,\pi)$, and for estimates close to $-\pi$ and π, the estimated distribution would become bimodal. Introducing k_i can avoid the bimodal issue as it unwraps $\hat{\theta}_i$ to values outside the range of $[-\pi,\pi)$ [29].

For all the $l\in(1,\cdots,\infty)$ components, μ_l is sampled from the conditional Gaussian distribution with mean χ_l, variance σ_l^2/ξ_l and concentration ξ_l, namely $(\mu_l|\sigma_l^2)\sim\boldsymbol{N}(\chi_l,\sigma_l^2/\xi_l)$, and σ_l^2 is sampled from an Inverse-Gamma distribution with shape η_l and scale γ_l, namely $\sigma_l^2\sim\boldsymbol{I-G}(\eta_l,\gamma_l)$. Thus, we can get the joint PDF of μ_l and σ_l^2 as follows

$$\begin{aligned} p(\mu_l,\sigma_l^2|\boldsymbol{\Psi}_l) \;&=\; p(\mu_l|\sigma_l^2,\chi_l,\xi_l)p(\sigma_l^2|\eta_l,\gamma_l) \\ &=\frac{\gamma_l^{\eta_l}e^{-\gamma_l/\sigma_l^2}}{\Gamma(\eta_l)\sigma_l^{2(\eta_l-1)}}\left(\frac{\xi_l}{2\pi\sigma_l^2}\right)^{\frac{1}{2}}e^{\left[-\frac{\xi_l}{2\sigma_l^2}(\mu_l-\chi_l)^2\right]} \\ &\propto\frac{1}{\sigma_l^{2\eta_l-1}}e^{\left\{-\frac{1}{2\sigma_l^2}[\xi_l(\mu_l-\chi_l)^2+2\gamma_l]\right\}}, \end{aligned} \qquad (19)$$

where $\boldsymbol{\Psi}_l=\{\xi_l,\eta_l,\chi_l,\gamma_l\}$ represents the hyperparameters of the IGMM.

Our aim is to estimate the class labels z_i for each $\hat{\theta}_i$, i.e. to which mixture component l we can assign $\hat{\theta}_i$. This is achieved by finding the maximum posterior probability of $z_i=l$, via iteratively updating the model parameters $\boldsymbol{\Psi}_l$ as discussed in the next section. For clarity, we show a graphical model for BNP in Fig. 4, and the notations in Table I.

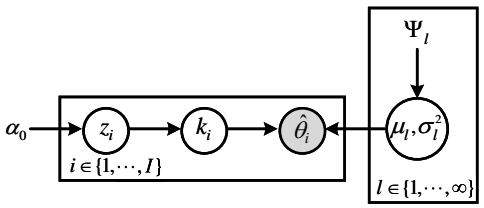


Fig. 4. Graphical model depicting the Bayesian nonparametrics.

B. Inference of Model Parameters

For model inference, we use the maximum a posteriori (MAP) [29] and Gibbs sampling [40] approaches. First, we compute the posterior probability of the shift k_i, given all other variables and all DOA measurements $\hat{\theta}_i$, and then determine the shift k_i that maximizes the posterior probability. Assuming a uniform prior for k_i, then the posterior probability follows a Student′s-t distribution $T_{2\eta_l^{(a)}}(\cdot)$, with degrees of freedom $2\eta_l^{(a)}$,

$$\begin{aligned} &p(k_i|\hat{\theta}_i,\hat{\boldsymbol{\theta}}_{\setminus i},z_i=l,\boldsymbol{k}_{\setminus i},\boldsymbol{z}_{\setminus i},\boldsymbol{\Psi}_l^{(a)}) \\ &\qquad\propto T_{2\eta_l^{(a)}}\left[(\hat{\theta}_i+2\pi k_i)\middle|\chi_l^{(a)},\frac{\gamma_l^{(a)}(\xi_l^{(a)}+1)}{\eta_l^{(a)}\xi_l^{(a)}}\right], \end{aligned} \qquad (20)$$

TABLE I
Notations

Symbol	Meaning	
$\hat{\Theta}$	DOA measurements at all the TF bins	$\hat{\Theta}=\{\hat{\theta}_1,\cdots,\hat{\theta}_i,\cdots,\hat{\theta}_I\}$
$\hat{\Theta}'$	DOA measurements at Q selected TF bins with highest weights	$\hat{\Theta}'=\{\hat{\theta}'_1,\cdots,\hat{\theta}'_q,\cdots,\hat{\theta}'_Q\}$
$\hat{\Theta}''$	Refined and re-aligned DOA measurements at the Q selected TF bins	$\hat{\Theta}''=\{\hat{\theta}''_1,\cdots,\hat{\theta}''_q,\cdots,\hat{\theta}''_Q\}$
l	Number of mixture components in IGMM	Number of speakers
z_i	Class label of the ith DOA estimate	
k_i	Shift of the ith DOA estimate	
$\boldsymbol{\Psi}_l$	Hyperparameters of IGMM	$\boldsymbol{\Psi}_l=\{\xi_l,\eta_l,\chi_l,\gamma_l\}$
$\boldsymbol{\Psi}_l^{(0)}$	Initial hyperparameters	$\boldsymbol{\Psi}_l^{(0)}=\{\xi_l^{(0)},\eta_l^{(0)},\chi_l^{(0)},\gamma_l^{(0)}\}$
$\boldsymbol{z}_{\setminus i}$	Set of class labels without the ith class	$\boldsymbol{z}_{\setminus i}=\{z_1,z_2,\cdots,z_{i-1}\}$
$\hat{\boldsymbol{\theta}}_{\setminus i}$	Set of all the DOA estimates without the ith DOA estimate	$\hat{\boldsymbol{\theta}}_{\setminus i}=\{\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_{i-1}\}$
$\boldsymbol{k}_{\setminus i}$	Set of all the shifts in the DOA estimates without the ith shift	$\boldsymbol{k}_{\setminus i}=\{k_1,k_2,\cdots,k_{i-1}\}$
a	BNP iteration number	Initial value $a=0$

where $\hat{\boldsymbol{\theta}}_{\setminus i}$ denotes the set of all the DOA measurements without $\hat{\theta}_i$, $\boldsymbol{k}_{\setminus i}$ is the set of all the shifts without k_i, $\boldsymbol{z}_{\setminus i}$ denotes the set of class labels without the ith label z_i, and $\boldsymbol{\Psi}_l^{(a)} = \{\xi_l^{(a)}, \eta_l^{(a)}, \chi_l^{(a)}, \gamma_l^{(a)}\}$ is the set of hyperparameters for component l, at the current iteration a.

Next, we calculate the posterior probability of $z_i = l$, i.e. the class label z_i belonging to the mixture component l, at the current iteration a, given all $\hat{\theta}_i$ and $\boldsymbol{\Psi}_l^{(a)}$. However, this probability cannot be computed directly. To address this issue, we use Gibbs sampling to approximate the posterior probability by drawing samples from two probability distributions that have simpler forms and can be computed, as follows

$$\begin{aligned} &p(z_i = l|\hat{\theta}_i, \hat{\boldsymbol{\theta}}_{\setminus i}, \boldsymbol{k}_{\setminus i}, \boldsymbol{z}_{\setminus i}, \boldsymbol{\Psi}_l^{(a)}) \\ &\quad \propto p(z_i = l|\boldsymbol{z}_{\setminus i}) p(\hat{\theta}_i|\hat{\boldsymbol{\theta}}_{\setminus i}, z_i = l, \boldsymbol{k}_{\setminus i}, \boldsymbol{z}_{\setminus i}, \boldsymbol{\Psi}_l^{(a)}), \end{aligned} \tag{21}$$

where the mixture component term $p(z_i = l|\boldsymbol{z}_{\setminus i})$ is updated by the Chinese Restaurant Process (CRP) as follows [30],

$$p(z_i = l|\boldsymbol{z}_{\setminus i}) = \begin{cases} \dfrac{i_l}{i + \alpha_0 - 1}, & \text{if } l \text{ is an existing mixture component} \\ \dfrac{\alpha_0}{i + \alpha_0 - 1}, & \text{if } l \text{ is a new mixture component} \end{cases} \tag{22}$$

where i_l is the number of DOA estimates assigned to the lth mixture component and the parameter α_0 is the concentration parameter of the Dirichlet Process [40].

The second term of (21) can be calculated by marginally integrating out the parameters $\{\mu_l, \sigma_l^2\}$ for the product of (18) and (19), and summing over k_i,

$$\begin{aligned} &p(\hat{\theta}_i|\hat{\boldsymbol{\theta}}_{\setminus i}, z_i = l, \boldsymbol{k}_{\setminus i}, \boldsymbol{z}_{\setminus i}, \boldsymbol{\Psi}_l^{(a)}) \\ &\propto \sum_{k_i} \int p(\hat{\theta}_i|\mu_l, \sigma_l^2, k_i) p(\mu_l, \sigma_l^2|\hat{\boldsymbol{\theta}}_{\setminus i}, z_i = l, \boldsymbol{k}_{\setminus i}, \boldsymbol{z}_{\setminus i}, \boldsymbol{\Psi}_l^{(a)})\, d\mu_l d\sigma_l^2 \\ &\propto \sum_{k_i} T_{2\eta_l^{(a)}} \left[\hat{\theta}_i + 2\pi k_i | \chi_l^{(a)}, \frac{\gamma_l^{(a)}(\xi_l^{(a)} + 1)}{\eta_l^{(a)} \xi_l^{(a)}} \right]. \end{aligned} \tag{23}$$

Note that, in the second line of (23), the integration cannot be computed directly. To address this issue, we have referred to the results in [41] in order to obtain the likelihood function shown in the third line of (23), which turns out to be also a Student's t-distribution.

After obtaining the posterior probabilities of the shifts k_i by (20) and class labels z_i by (21), we can recalculate the parameters $\boldsymbol{\Psi}_l^{(a+1)}$ at the next iteration $(a+1)$ according to $\hat{\theta}_i$. Here, we adopt the theory of conjugate priors in the Classical Bayesian Statistics [40], namely, the posterior probability and the prior probability belong to the same distribution family, and derive the update formula for the model parameters (details given in Appendix A), as follows

$$\xi_l^{(a+1)} = \xi_l^{(a)} + 1, \tag{24}$$

$$\eta_l^{(a+1)} = \eta_l^{(a)} + \frac{1}{2}, \tag{25}$$

$$\chi_l^{(a+1)} = \frac{\xi_l^{(a)} \chi_l^{(a)} + (\hat{\theta}_i + 2\pi k_i)}{\xi_l^{(a)} + 1}, \tag{26}$$

$$\gamma_l^{(a+1)} = \gamma_l^{(a)} + \frac{\xi_l^{(a)} (\hat{\theta}_i + 2\pi k_i - \chi_l^{(a)})^2}{2(\xi_l^{(a)} + 1)}, \tag{27}$$

where a is the BNP iteration number with initial value $a = 0$.

In summary, by using (20) and (21), we can obtain the posterior probability of k_i and z_i for each DOA measurement $\hat{\theta}_i$, thereby assign this $\hat{\theta}_i$ to the corresponding mixture component l (i.e., $z_i = l$) that gives the highest probability among all the components. Then, we recalculate the hyperparameters using (24), (25), (26) and (27) to update the model. This process is iterated until all $\hat{\theta}_i$s in $\hat{\Theta}$ have been assigned to one of the l mixture components (or clusters). Note that l is also updated in each iteration. Finally, the number of speakers D and the DOAs $\hat{\theta}_d$, $d = 1, ..., D$, are obtained as the number of mixture components l and the cluster centroids in the final iteration, upon convergence of the BNP iterations.

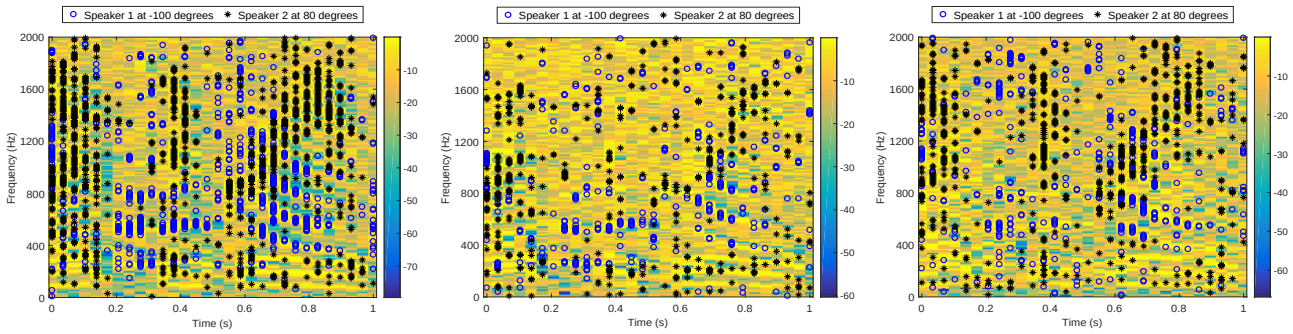
VI. Robust Parameter Estimation

Although the IGMM can adjust the model adaptively to locate multi-speakers on the basis of DOAs, the permutations of the estimates are usually random, and not every estimate is valid, in other words, the established mixture model may be easily affected by the random permutation and the invalid or erroneous estimates, resulting in the inaccurate localization results. In addition, due to the presence of reverberation and background noise in an indoor environment, some DOA estimates may be inaccurate, and are likely to spread over the whole possible DOA region, which can lead to inaccurate localization. To address this problem, we propose a reliable TF bin selection and permutation alignment scheme on the basis of the mixture weights.

A. Calculation of Mixture Weights

Speech signals are sparse in the TF domain, and some TF bins may contain only background noise. Moreover, due to the presence of room reverberation, the received signals by sensors usually vary over different TF bins. As a result, only a fraction of TF bins corresponds to the accurate source DOA in the estimated instantaneous DOAs. Therefore, we develop a scheme for selecting reliable TF bins to improve the robustness of the proposed algorithm against noise and room reverberation, in terms of the mixture weights derived from the following features.

- Power: The power at each TF bin can be expressed as $E(k,t) = \mid P_m(k,t) \mid^2$. According to [24], we know that higher power at a TF bin is less affected by the environmental noise than lower power. Thus, this feature can be used to mitigate noise.
- Power Ratio: The power ratio at a TF bin can be obtained from the two TF bins in the adjacent time frames and expressed as $\frac{E(k,t)}{E(k,t-1)}$ [43]. If the ratio is greater than 1, then the energy of the current time frame is increased



(a) (b) (c)

Fig. 5. The locations of reliable TF bins in the spectrogram of the two speakers. One (speaker 1) is at $\theta_1 = -100^\circ$ and the other (speaker 2) is at $\theta_2 = 80^\circ$. (a) $RT_{60} = 200$ ms and SNR $= 20$ dB (RMSE $= 1.27^\circ/1.58^\circ$ with/without TF selection); (b) $RT_{60} = 200$ ms and SNR $= 5$ dB (RMSE $= 3.21^\circ/8.06^\circ$ with/without TF selection); (c) $RT_{60} = 500$ ms and SNR $= 20$ dB (RMSE $= 3.72^\circ/7.52^\circ$ with/without TF selection).

compared with the previous time frame, which may suggest that the speech signal transitions from an unvoiced part to a voiced part. These TF bins have relatively little influence on room reverberation, which can be used as an important parameter to represent the TF bin reliability.

- Local Variance: The reliable instantaneous DOAs often exhibit low local DOA variances [44]. In contrast, high variances, however, indicate the TF regions where the instantaneous DOA estimates are corrupted by the room reverberation. In other words, the TF bins with low local DOA variance, are less affected by noise and room reverberation. The local DOA variance at the TF bin can be expressed as

$$\sigma^2_{\hat{\theta}}(k,t) = \frac{1}{N_{tf}-1} \sum_{(k,t)\in\Omega_{(k,t)}} [\hat{\theta}(k,t) - \bar{\theta}(k,t)]^2, \quad (28)$$

where $\Omega_{(k,t)}$ denotes the neighborhood centered at (k,t), N_{tf} is the number of TF bins within $\Omega_{(k,t)}$, and $\bar{\theta}(k,t) = \frac{1}{N_{tf}} \sum_{(k,t)\in\Omega_{(k,t)}} \hat{\theta}(k,t)$ is the local DOA mean at (k,t).

With the power, the power ratio and the local variance, we can now develop a scheme for the selection of reliable TF bins using a mixture weight $W(k,t)$, computed as

$$W(k,t) = 3[w_p(k,t) w_{pr}(k,t) w_{var}(k,t)]^2, \quad (29)$$

where $w_p(k,t)$ denotes the weight of power at the TF bin (k,t), $w_{pr}(k,t)$ denotes the weight of power ratio, and $w_{var}(k,t)$ denotes the weight of local DOAs variance, determined via the sigmoid compression, respectively

$$w_p(k,t) = \frac{1}{1+e^{-\alpha_1[\log E(k,t) - \beta_1]}}, \quad (30)$$

$$w_{pr}(k,t) = \frac{1}{1+e^{-\alpha_2[\log \frac{E(k,t)}{E(k,t-1)} - \beta_2]}}, \quad (31)$$

$$w_{var}(k,t) = \frac{1}{1+e^{-\alpha_3[\log \sigma^2_{\hat{\theta}}(k,t) - \beta_3]}}. \quad (32)$$

The $\alpha_1, \alpha_2, \alpha_3$ and the $\beta_1, \beta_2, \beta_3$ are the sigmoid slope and center parameters, respectively [45]. We set them empirically in our experiments as discussed in Section VII-C. In addition, the purpose of using a square and a constant factor 3 in (29) is to empirically enlarge the difference among different weights which can improve the performance as observed in our experiments.

An example is given in Fig. 5, where we plot $B_{\mathrm{AH}}(k\check{r}, t, \theta)$ obtained from (16) with the color showing the values of $10\log_{10}(B_{\mathrm{AH}})$, to demonstrate the TF selection based on $W(k,t)$, where the reliable TF bins are indicated, for which the difference between the estimated and true DOAs is less than 5 degrees. Here, the mixtures of two speech sources, at -100°, and 80°, respectively, are considered, and the reverberation times RT_{60} are {200 ms, 500 ms} and the SNR are {5 dB, 20 dB}. From Fig. 5, it is noticed that a certain amount of reliable DOA estimates were selected and demonstrated in the spectrogram under different environments (e.g. low SNR or high reverberation), which further confirms the validity of the proposed method for the TF bin selection based on the mixture weights. The improved performance by the TF selection can be seen from the RMSE results given on the caption of Fig. 5.

B. Refinement of DOA Measurements

The mixture weights can be used jointly to determine the reliability of the TF bins to improve the robustness against noise and room reverberation. Note that the greater the mixture weight is, the higher the reliability of the TF bin is, and vice versa. Therefore, we select Q DOA estimates with the highest weight $\hat{\Theta}' = \{\hat{\theta}'_1, \cdots, \hat{\theta}'_q, \cdots, \hat{\theta}'_Q\}$, and only use these data in the permutation procedure, as described below. The value of Q is given in Section VII-C.

(i) We form a histogram from $\hat{\Theta}'$ and smooth it by applying an averaging filter with a window. Denote each bin of the smoothed histogram as υ, then its cardinality, $y(\upsilon)$, is given by

$$y(\upsilon) = \sum_{q=1}^{Q} \omega\left(\frac{\upsilon \times 360^\circ/\Upsilon - \hat{\theta}'_q}{h_Q}\right), 1 \le \upsilon \le \Upsilon, \quad (33)$$

where Υ is the number of bins in the histogram, and $\omega(\cdot)$ is the blackman window of length h_Q, where empirically $h_Q = 21$ [31].

(ii) We normalize the smoothed histogram $y(\upsilon)$ by

$$y'(\upsilon) = \frac{y(\upsilon)}{\max\{y(\upsilon)\}}. \quad (34)$$

(iii) We perform peak search of $y'(\upsilon)$ to find all the peaks that exceed a pre-defined threshold th_{ps}, and obtain the pre-estimation $\hat{\theta}_{pk} = \{\hat{\theta}_{pk1}, \hat{\theta}_{pk2}, \cdots\}$. The choice of th_{ps} is discussed in Section VII-C.

(iv) We calculate the difference between the qth DOA estimation and the pre-estimation, and assign the minimum difference to diffm_q as

$$\mathrm{diffm}_q = \min\{\hat{\theta}'_q - \hat{\theta}_{pk}\}. \quad (35)$$

(v) By sorting diffm_q in ascending order, we realign the DOAs estimation $\hat{\Theta}'$ and obtain the new $\hat{\Theta}'' = \{\hat{\theta}''_1, \cdots, \hat{\theta}''_q, \cdots, \hat{\theta}''_Q\}$, which is used as the input to the BNP algorithm discussed in Section V.

C. Parameter Update with Mixture Weights

We update the hyper-parameters in terms of the mixture weights. With the parameter update, the greater the mixture weights of the TF bins are, the stronger the decisive role and the update degree are. Incorporating the mixture weights as in (28), we can obtain the new update formula [41]:

$$\xi_l^{(a+1)} = \xi_l^{(a)} + W_q, \quad (36)$$

$$\eta_l^{(a+1)} = \eta_l^{(a)} + \frac{1}{2}W_q, \quad (37)$$

$$\chi_l^{(a+1)} = \frac{1}{\xi_l^{(a+1)}}\left[\xi_l^{(a)}\chi_l^{(a)} + W_q(\hat{\theta}''_q + 2\pi k_q)\right], \quad (38)$$

$$\gamma_l^{(a+1)} = \gamma_l^{(a)} + \frac{1}{2}\Big[W_q(\hat{\theta}''_q + 2\pi k_q)^2 + \xi_l^{(a)}(\chi_l^{(a)})^2 - \xi_l^{(a+1)}(\chi_l^{(a+1)})^2\Big], \quad (39)$$

In addition, W_l^{sum} denotes the sum of the mixture weights for the lth class of the DOA estimation. The number of W_l^{sum} that is greater than the threshold th_W is taken as the estimated number of speakers. The threshold th_W can be calculated as

$$th_W = b_0\left[mean(W_l^{sum}) + \sqrt{var(W_l^{sum})}\right], \quad (40)$$

where $mean(\cdot)$ and $var(\cdot)$ represent taking the mean and variance over its argument, respectively, and b_0 is a constant.

Finally, the proposed TF-MX-BNP-AHB method is summarized in Algorithm 1. Note that, the method starts with no DOAs estimation assigned to a mixture component and no mixture components created. The DOA estimates are assigned to the chosen or created mixture components in the first iteration and then added to the corresponding statistics. The hyperparameters obtained after the final iteration are then used for the estimation of the DOAs of the speakers. The lines 3 and 4 of Algorithm 1 are corresponding to stage 1 described in Section III, the lines 7, 8 and 9 are corresponding to stage 2 (here, for line 8, if the robust parameter estimation method is not used, then the hyperparameters Ψ_l are updated using equations (24), (25), (26) and (27)), and the lines 5 and 6 are corresponding to stage 3. The algorithm stops finding new DOA clusters when it has gone through all the obtained DOA estimates in $\hat{\Theta}''$.

Algorithm 1 The proposed TF-MX-BNP-AHB method

Input: Microphone signals $p_m(\tau)$, for $m = 1$ to M
Output: Number of speakers D and their DOAs $\hat{\theta}_d$
1: **Initialize** M, N, r, $\breve{r}$, $\xi_l^{(0)}$, $\eta_l^{(0)}$, α_0, b_0.
2: **procedure**
3: Perform STFT on $p_m(\tau)$ to obtain $P_m(k,t)$;
4: Perform TF-AHB to obtain DOA measurements $\hat{\Theta}$ using (16) and (17);
5: Construct mixture weights $W(k,t)$ by (29), (30), (31) and (32) to select reliable TF bins and obtain Q DOAs estimates $\hat{\Theta}'$ with the highest weights;
6: Calculate diffm_q by (33), (34) and (35), and sort them in ascending order. Realign $\hat{\Theta}'$ and obtain the new $\hat{\Theta}''$;
7: Input $\hat{\Theta}''$ to the IGMM, and calculate the posterior probability of shift k_q and class label z_q of current DOA estimate by (20) and (21);
8: Update the hyperparameters $\mathbf{\Psi}_l$ of corresponding class with mixture weights by (36), (37), (38) and (39);
9: Input the updated hyperparameters $\mathbf{\Psi}_l$ to the IGMM. Cluster and obtain the estimated number of speakers D and their DOAs $\hat{\theta}_d$.
10: **end procedure**

VII. Experimental Evaluations

This section studies the performance of the proposed method through simulations and real data experiments. DOA estimation using the proposed method is investigated and compared to baseline methods under different number of sources and acoustic environments. The section starts with a description of datasets, evaluation metrics and parameter setup, and then presents the experimental results.

A. Datasets

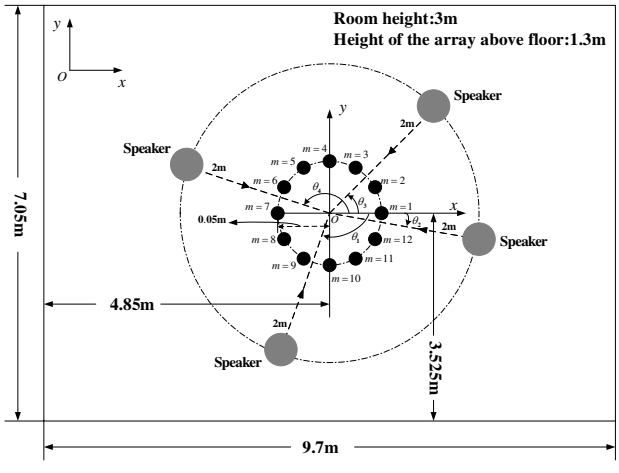


Fig. 6. Illustration of the simulation setup. The black solid dots, distributed uniformly on the dash-dotted circle with a radius of 0.05 m, denote the UCA. The gray circles, distributed around the dash-dotted circle with a radius of 2 m, denote the speakers.

Both simulated data and real-world data are generated. The simulation setup is shown in Fig. 6. The dimensions of the

simulated rectangular room are 9.7 m × 7.05 m × 3 m. To generate the RIRs [46] from speaker sources to sensors, we use a software that is based on the well-known image method for simulating a reverberant environment in a room [47]. The UCA with M = 12 equidistant omnidirectional sensors and the radius of r = 0.05 m (the radius of virtual array $\breve{r} = 2r$) is placed in the center of the room at $(4.85, 3.525, 1.3)$ m, coinciding with the origin of the x and y axes. The speakers are located at the same height as the microphone array with distance from the speaker to the center of the array being 2 m. The additive noises on the sensors are mutually uncorrelated white Gaussian, and also are uncorrelated with the speech signal. Several levels of room reverberation with various reverberation times are tested, which will be specified later. The sound speed is 340 m/s. Speech signals of 2 s length, sampled at 16 kHz, are chosen randomly from the well-known TIMIT speech database [48]. 100 Monte Carlo simulations are conducted in each trial, and the source was convolved with the simulated RIRs from the source to every microphone. For all the evaluated algorithms, the STFT is calculated using a Hamming window of 1024 samples with 50% overlap between consecutive frames.

In order to analyze the source localization performance comprehensively, we consider four different aspects in the simulations:

(1) *Angular Distance*: To investigate the spatial resolution for our proposed method, we consider various angular distances for pairs of speakers.

(2) *Different Types of Noise*: To evaluate the impact of noise types, we consider five different types of background noise, in addition to white Gaussian noise.

(3) *Effect of Room Reverberation and Additive Noise*: To evaluate the influence of room reverberation and additive noise on the performance of our proposed method, we consider three different multi-speaker scenarios under varying levels of noise and reverberation times. Herein, the reverberation time RT_{60} is varied from 200 to 700 ms with a step increase of 100 ms and the SNR is varied from 0 to 20 dB with a step increase of 5 dB. These three scenarios are detailed below.

- Two sources: Two different speakers (Speaker 1, female; Speaker 2, male) are used and they are placed at DOAs of $-30°$ (Speaker 1) and $0°$ (Speaker 2), respectively.
- Three sources: Three different speakers (Speaker 1, female; Speaker 2, male; Speaker 3, female) are used and they are placed at DOAs of $-70°$ (Speaker 1), $0°$ (Speaker 2) and $45°$ (Speaker 3), respectively.
- Four sources: Four different speakers (Speaker 1, female; Speaker 2, male; Speaker 3, female; Speaker 4, male) are used and the speakers are located at DOAs of $-110°$ (Speaker 1), $0°$ (Speaker 2), $30°$ (Speaker 3) and $100°$ (Speaker 4), respectively.

(4) *Effect of Different Positions*: To assess the impact of different positions on our proposed method, we consider diverse locations for two, three and four sources, respectively, and demonstrate their performance.

To further evaluate the effectiveness of the proposed method, we also recorded speech in a real rectangular conference room with dimensions of approximately 9.7 m × 7.05 m × 3 m and a reverberation time of 350 ms. A small-sized array was placed horizontally around the center of the room, and the other conditions resembled those in the above simulations. A photograph of the microphone array is shown in Fig. 7.

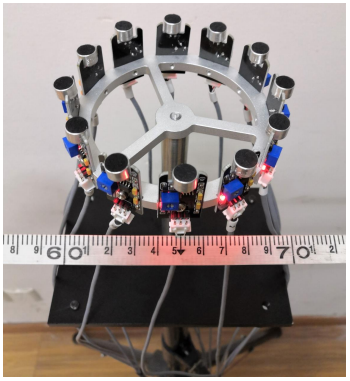


Fig. 7. Photograph showing the uniform circular sensor array used in the experiments. The radius of the circular microphone array is 0.05 m.

The sensors used in the real-world experiments were Arduino omnidirectional sensors with working voltage 5 V. The received sensor signals were sampled with a sampling frequency of 16 kHz through a data-acquisition device (NI-USB-6363; National Instruments) with 16-bit resolution. In the real-experiments, the actual speaker locations of the sound sources were determined using protractors and rulers, and 50 independent experiments were conducted for each trial. The duration of each of the received sensor signals was 2 s.

B. Evaluation Metrics

To facilitate evaluations, we use the root-mean-square error (RMSE) and the source detecting success rate (SDSR) as performance metrics, which are defined as:

$$RMSE = \sqrt{\frac{1}{DO}\sum_{d=1}^{D}\sum_{o=1}^{O}\left[\hat{\theta}_d(o) - \theta_d\right]^2}, \quad (41)$$

where $\hat{\theta}_d(o)$ is the DOA estimate of the dth speakers DOA θ_d for the oth simulation or real-world experiment and O is the number of simulations or real-world experiments.

$$Source\ Detecting\ Success\ Rate = \frac{\hat{O}_{ds}}{O} \times 100\%, \quad (42)$$

where $\hat{O}_{ds}$ is the number of experiments with source detecting success, i.e., the error between the estimated and true DOAs is no greater than $15°$ [22].

C. Parameters Setup

Over a number of trials, the corresponding parameters of the proposed method are set empirically to $\xi_l^{(0)} = 0.01$, $\eta_l^{(0)} = 0.01$, $\alpha_0 = 150$, $b_0 = 0.5$, $\chi_0 = \frac{\sum_{q=1}^{Q} W_q \hat{\theta}_q''}{\sum_{q=1}^{Q} W_q}$, and $\gamma_0 = \frac{\sum_{q=1}^{Q} W_q (\hat{\theta}_q'' - \chi_l^{(0)})^2}{\sum_{q=1}^{Q} W_q}$. We set Q to be equal to 20 % of the total number of TF bins.

We set the parameters in (30), (31), and (32) with empirical tests based on our data. Here, we take the selection of α_1 and β_1 as an example. As we know, the value of α_1 controls the

orientation of the S-shape of the sigmoid function, and we want the weight $w_p(k,t)$ to be increasing with the increase in value of $E(k,t)$. We have tested $\alpha_1 \in [-10, 10]$ and found that $\alpha_1 \in [-10, -1]$ meets our requirement. Thus, we choose $\alpha_1 = -1$ in our experiments. The parameter β_1 controls the value of weight $w_p(k,t)$. In order to select the TF bins with high power, we want the weights to be greater 0.5. In practice, however, if the value of the weight is too large (e.g., 0.9), we may lose some useful TF bins and as a result, we may not have sufficient TF bins for the localization of the sources. Thus, we consider a trade-off and select $w_p(k,t)$ to be around 0.5, which corresponds to $\beta_1 = -8.5$. Other parameters α_2, β_2, α_3 and β_3 are chosen in a similar way, as $\alpha_2 = -6, \beta_2 = -0.5$; and $\alpha_3 = 3, \beta_3 = 8$. The parameter values may not be optimal for other data, which may need to be re-tuned similarly.

For the pre-defined threshold th_{ps} discussed in Section VI-B for peak finding, we tested $th_{ps} \in [0.25, 0.35]$. Although the values of several peaks are lower than 0.3, we empirically choose the value of threshold $th_{ps} = 0.3$, which seems to be appropriate for most scenarios.

D. Baseline Methods

The performance of the proposed TF-MW-BNP-AHB is evaluated and compared with several baselines including the TF-AHB, TF-CHB [20], CH-SS-PIV [22] and DPD-MUSIC [23] methods in both simulated and real room environments.

TF-AHB: This is the AHB method over a TF processing framework where DOAs are estimated by taking the histogram with maximum power at the corresponding arrival direction, as aforementioned, using (15) and (16).

TF-CHB: This is a broadband source localization method based on combination of TF processing and CHB. The histogram of DOA estimates computed over the TF plane shows clear peaks corresponding to the locations of different sources by taking the direction of maximum CHB output power.

CH-SS-PIV: This method performs subspace (SS) decomposition of the spatial covariance matrix and computes a PIV in the directions of the sound sources to directly obtain the DOA estimates in the circular harmonic domain.

DPD-MUSIC: In this method, the MUSIC algorithm is used with the DPD test. This test aims to estimate DOA of the sources by identifying the TF bins in the microphone signal that contain contributions from only one significant source and no significant contribution from room reflections.

Here we assume that the number of speakers is known *a priori* in all the compared methods except for our proposed method.

E. Source Localization Results in Simulation Experiments

1) Angular Distance: In our first set of simulations we investigated the spatial resolution of our proposed method, i.e., how close two sources can be in terms of angular distance while accurately estimating their DOAs.

Fig. 8 shows the RMSE when the SNR is varied from 0 to 20 dB with a step increase of 5 dB and the reverberation time of $RT_{60} = 300$ ms, for pairs of active speakers with angular separations from 180° down to 20°. The simulation result shows clearly that our method performs well for most angular separations. For example, when the separation is around 30°, the RMSE is about 5° for SNR = 0 dB, which shows that the proposed method has a good angular distance in source localization.

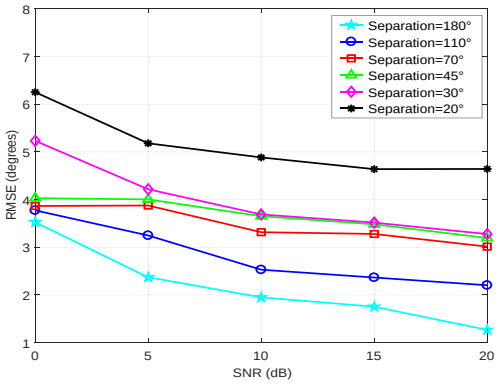


Fig. 8. RMSE of DOA estimation versus pairs of active speakers with angular separations from 180° down to 20° and $RT_{60} = 300$ ms.

2) Different Types of Noise: This experiment was carried out to evaluate the performance of proposed method in the presence of different types of noise. For contrasting, five types of noise (i.e., Destroyerops, Volvo, Factory1, Babble and Buccaneer1) from the Noisex-92 dataset [49] were used as background noise sources. We also compared them with white Gaussian noise (WNG). Here, we consider a pair of active speakers with angular separations 70°.

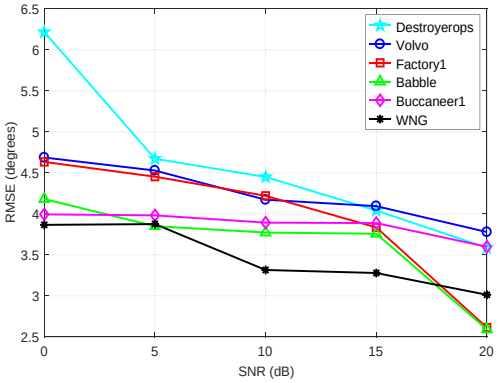
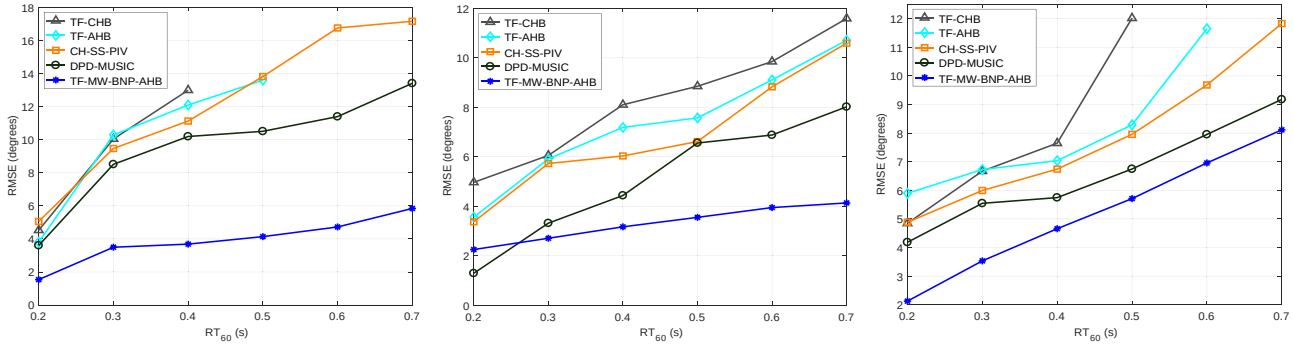


Fig. 9. RMSE of DOA estimation with different types of noise and $RT_{60} = 300$ ms.

Fig. 9 shows the RMSE of DOA estimation with different types of noise when the SNR is varied from 0 to 20 dB with a step increase of 5 dB and the reverberation time of $RT_{60} = 300$ ms. The simulation results show that our method performs well for the five types of noises tested. The RMSE is mostly smaller than 5°, except for the case of Destroyerops noise with SNR at 0 dB.

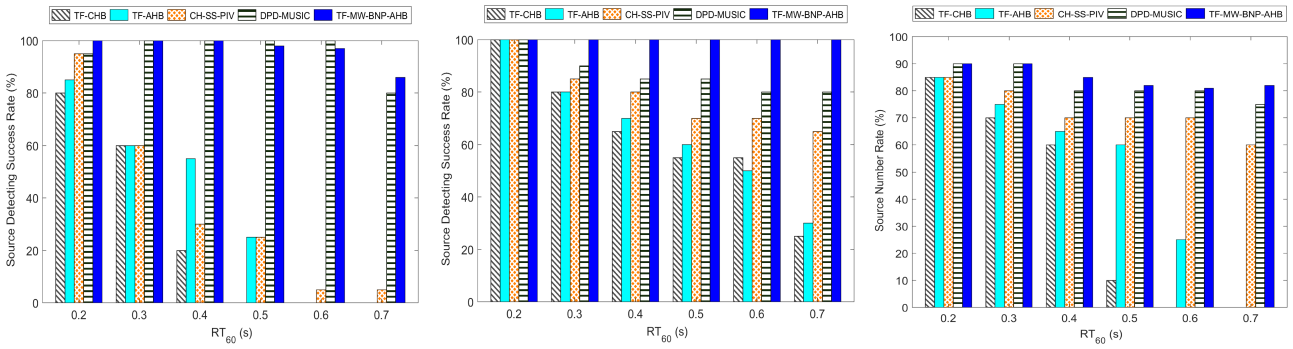
3) Effect of Room Reverberation and Additive Noise: Fig. 10 and Fig. 11 show the RMSE and the Source Detecting Success Rate of each method when the reverberation time RT_{60} is varied from 200 to 700 ms with a step increase of 100 ms and the level of noise in terms of SNR is 15 dB.

From Fig. 10, we can see that the proposed TF-MW-BNP-AHB method is quite robust to the change in reverberation for



(a) (b) (c)

Fig. 10. Effects of room reverberation on the performance of each method for RMSE with SNR $= 15$ dB: (a) Two sources; (b) Three sources; (c) Four sources.



(a) (b) (c)

Fig. 11. Effects of room reverberation on the performance of each method for Source Detecting Success Rate with SNR $= 15$ dB: (a) Two sources; (b) Three sources; (c) Four sources.

three scenarios. The RMSE for four sources is slightly higher than those for two and three sources, due to the increase in the number of sources. The DPD-MUSIC method performs considerably worse than the proposed method, except for the case e.g. RT_{60} = 200 ms for three sources. This may be because the DPD test model described in [23] computes the spatial spectrum over TF regions where one direct path is dominant which may reduce the number of TF bins that passed the DPD test [2]. This method is also relatively robust to reverberation changes. The performance of CH-SS-PIV has a similar trend to that of the DPD-MUSIC at the low to modest level of reverberation, but degrades more considerably at the higher level reverberation. This may be because of improved DOA estimates given by the subspace technique for the eigenbeams in CH-SS-PIV. Nevertheless, the increase in RT_{60} has a negative influence on the signal subspace, thus the localization accuracy of the sources. In contrast, the TF-CHB and TF-AHB perform poorly in three scenarios, especially under higher reverberation conditions, and have failed to localize the sources when $RT_{60} = 600$ ms or 700 ms. The main reason is that in two and four sources scenarios, when the adjacent sources are separated by a smaller angle, e.g. 30°, some adjacent sources may be partially overlapped, due to the presence of strong reverberations, which results in significantly degraded localization results. It was reported, however, that both methods worked well for the sources which are widely separated by more than 45° [20].

Note that, CH-SS-PIV and DPD-MUSIC perform better in the case of more sources (e.g. four sources) than that for two sources at high reverberation. This is mainly because the angular distances set for the case of two sources (Fig. 10(a)) are smaller than those of three sources (Fig. 10(b)) and four sources (Fig. 10(c)). We also tested all the compared methods for four sources with the same angular distance as used in the case of two sources (e.g. 30°), and their performance is consistent with the case of two sources, which decreases especially at high reverberation. Such results were not included here for space limitation.

From Fig. 11, we can see that the Source Detecting Success Rate also degrades with the increase in reverberation levels and show the similar performance trend as in RMSE, as expected. Our method provides more accurate estimates of the number of sources than the baseline methods DPD-MUSIC, CH-SS-PIV, TF-AHB and TF-CHB, while the latter two perform poorly when the level of reverberation is high (e.g. $RT_{60} = 600$ ms or 700 ms). Note here that assuming a known number of speakers gives all baseline methods (apart from our proposed) a considerable advantage, as this avoids the errors due to the estimation of the number of speakers.

Fig. 12 and Fig. 13 show the RMSE and the Source Detecting Success Rate of each method when the SNR is varied from 0 to 20 dB with a step increase of 5 dB with the reverberation time of $RT_{60} = 300$ ms.

From Fig. 12, we can see that the performance of the proposed method is less affected by the changes in the input SNR for the three scenarios, as compared with the baseline

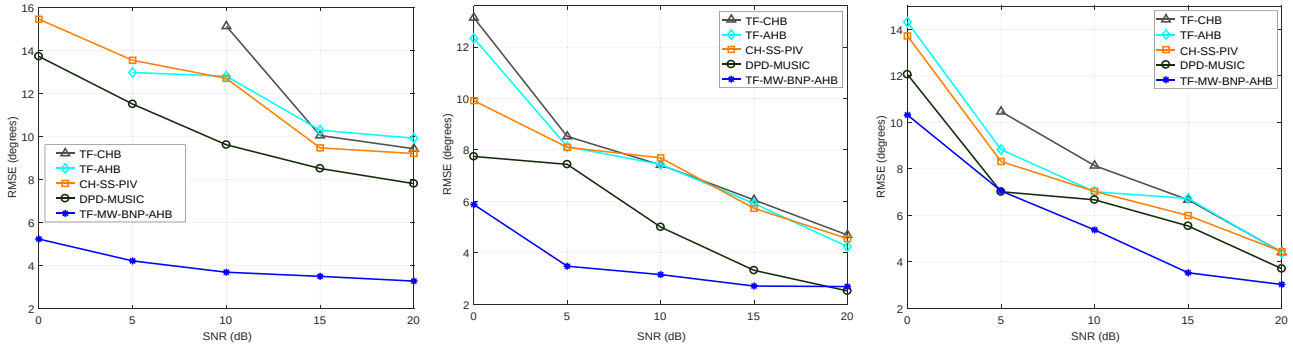


Fig. 12. Effects of additive noise on the performance of each method for RMSE with $RT_{60} = 300$ ms: (a) Two sources; (b) Three sources; (c) Four sources.

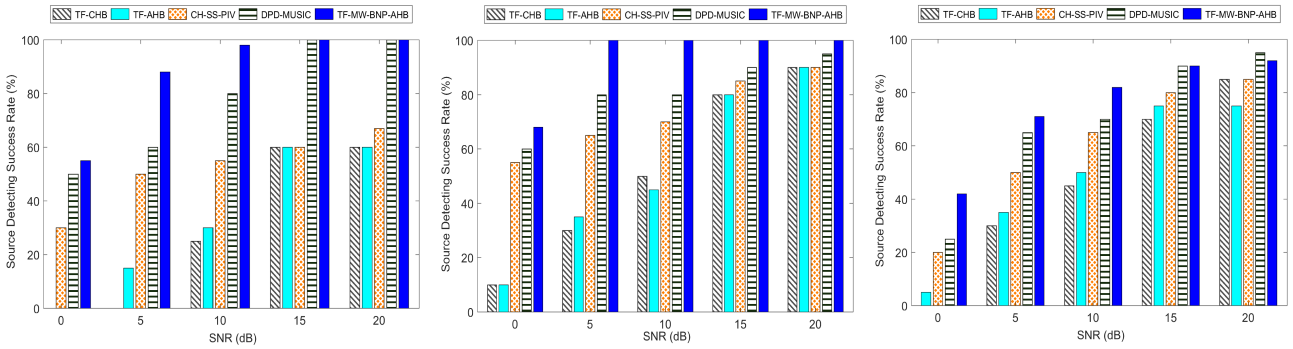


Fig. 13. Effects of additive noise on the performance of each method for Source Detecting Success Rate with $RT_{60} = 300$ ms: (a) Two sources; (b) Three sources; (c) Four sources.

methods. The simulation results show that the RMSE of the proposed method is less than 10° when $RT_{60} = 300$ ms, for most SNR levels. Both DPD-MUSIC and CH-SS-PIV show a similar trend, but with higher RMSE, as compared with the proposed method. Although both methods use subspace decomposition to suppress noise, DPD-MUSIC achieves better accuracy than CH-SS-PIV as DPD-MUSIC considers the TF regions with a single dominant source [22].

However, regarding the robustness, the TF-CHB and the TF-AHB methods show higher variances over noisy conditions. They produce reliable DOA estimates at low reverberation or high SNR, but have higher RMSE when the acoustic conditions deteriorate (e.g. the RMSE result at SNR = 0 dB was even invalid, therefore not shown in the figure). One reason is that, under a high level of noise, there may be mis-detection or spurious estimates, especially in cases with adjacent sources of lower separation as described before. The other one is that these two methods use higher mode order, which may amplify the influence of the noise interference [20], [26], and hence impact adversely on their ability to localize the sources correctly. In addition, from Fig. 13, it is clear that in all the tested cases, the proposed method can reliably detect source number with over 70% in most noise conditions, except when SNR= 0 dB, the correct rate is 55%, 68% and 42% for two, three and four sources, respectively. Again assuming a known number of speakers (which avoids the errors due to estimation of the number of speakers at adverse conditions), the DPD-MUSIC and CH-SS-PIV are less accurate than the proposed method in terms of Source Detecting Success Rate, while the TF-CHB and TF-AHB may not find the correct numbers of sources as the noise level increases.

In general, compared with the baseline methods, the proposed TF-MW-BNP-AHB offers higher accuracy in DOA estimation in reverberant environments and gives also fairly consistent results over different noise levels.

4) Effect of Different Source Positions: The RMSEs of the compared methods for two sources, three sources and four sources are given in Table II, Table III and Table IV respectively, when the reverberant time RT_{60} = 300 ms and SNR = 10 dB. From these tables we can see that the proposed method performs the better than the baselines in terms of the RMSE for different speakers located at a variety of positions in the room. In comparison, simply picking the maximum from the DOA histogram such as in TF-CHB, does not work well, as it gives an overall larger RMSE in these conditions.

F. Source Localization Results in Real-World Experiments

Table V, Table VI and Table VII show the RMSE of all the five evaluated methods for multi-speaker location at different positions in real-world experiments (as in the simulations study). As can be seen, the localization results behave in a similar manner to those found in the simulation results above. The proposed method manages to localize all the speakers with the least RMSEs in the real environment, which imply that our method outperforms the four baseline methods. In addition, from these tables we can also see that the performance of the

TABLE II
TWO SOURCES: THE RMSE OF THE EVALUATED METHODS

True sources	TF-CHB	TF-AHB	CH-SS-PIV	DPD-MUSIC	TF-MW-BNP-AHB
$(-30°, 0°)$	$15.13°$	$12.81°$	$12.70°$	$9.62°$	$3.69°$
$(-35°, 10°)$	$14.56°$	$9.61°$	$10.30°$	$7.21°$	$2.76°$
$(-50°, 20°)$	$13.00°$	$7.52°$	$8.23°$	$7.62°$	$2.33°$
$(-80°, 30°)$	$11.40°$	$5.09°$	$5.16°$	$3.81°$	$1.84°$
$(-100°, 80°)$	$5.38°$	$4.47°$	$3.94°$	$1.58°$	$0.86°$

TABLE III
THREE SOURCES: THE RMSE OF THE EVALUATED METHODS

True sources	TF-CHB	TF-AHB	CH-SS-PIV	DPD-MUSIC	TF-MW-BNP-AHB
$(-45°, 0°, 30°)$	$7.87°$	$7.77°$	$9.33°$	$5.80°$	$4.37°$
$(-70°, 0°\ 45°)$	$7.42°$	$7.44°$	$7.69°$	$5.00°$	$3.15°$
$(-120°, -10°, 60°)$	$5.72°$	$5.45°$	$5.44°$	$4.36°$	$2.96°$
$(-150°, 30°, 100°)$	$4.43°$	$3.79°$	$3.98°$	$3.41°$	$2.37°$

TABLE IV
FOUR SOURCES: THE RMSE OF THE EVALUATED METHODS

True sources	TF-CHB	TF-AHB	CH-SS-PIV	DPD-MUSIC	TF-MW-BNP-AHB
$(-30°, 0°, 45°, 115°)$	$10.10°$	$8.92°$	$8.31°$	$8.17°$	$6.50°$
$(-110°, 0°\ 30°, 100°)$	$8.14°$	$7.02°$	$7.03°$	$6.67°$	$5.38°$
$(-150°, -40°, 30°, 120°)$	$6.71°$	$7.00°$	$5.93°$	$4.72°$	$4.24°$
$(-160°, -70°, 20°, 110°)$	$4.77°$	$5.34°$	$4.19°$	$3.43°$	$3.40°$

proposed TF-MW-BNP-AHB method is less affected by practical reverberant and noisy environments, whose results have less variations, when compared with the baseline methods.

VIII. CONCLUSION

We have presented an indoor multi-speaker localization method TF-MW-BNP-AHB. This method exploits the acoustic holography beamforming technique in the TF domain which is less restricted by the practical constraints, such as the array shape and the number of sensors, as compared with conventional circular harmonic DOA estimation methods.

We have developed a BNP method based on the AHB model for multi-source localization without the knowledge of the number of sources, and the use of mixture weighting

TABLE V
TWO SOURCES: THE RMSE OF THE EVALUATED METHODS IN REAL-WORLD EXPERIMENTS

True sources	TF-CHB	TF-AHB	CH-SS-PIV	DPD-MUSIC	TF-MW-BNP-AHB
$(-30°, 0°)$	$15.50°$	$9.50°$	$10.41°$	$6.50°$	$5.37°$
$(-35°, 10°)$	$14.00°$	$8.50°$	$7.29°$	$6.00°$	$4.59°$
$(-50°, 20°)$	$10.50°$	$8.00°$	$7.07°$	$4.50°$	$3.82°$
$(-80°, 30°)$	$11.50°$	$7.00°$	$7.16°$	$3.50°$	$3.34°$
$(-100°, 80°)$	$7.00°$	$6.50°$	$3.65°$	$2.50°$	$1.22°$

TABLE VI
THREE SOURCES: THE RMSE OF THE EVALUATED METHODS IN REAL-WORLD EXPERIMENTS

True sources	TF-CHB	TF-AHB	CH-SS-PIV	DPD-MUSIC	TF-MW-BNP-AHB
$(-45°, 0°, 30°)$	$11.67°$	$12.00°$	$9.61°$	$8.67°$	$6.93°$
$(-70°, 0°\ 45°)$	$9.66°$	$10.67°$	$9.04°$	$7.33°$	$6.82°$
$(-120°, -10°, 60°)$	$8.33°$	$7.33°$	$6.78°$	$4.33°$	$3.75°$
$(-150°, 30°, 100°)$	$8.67°$	$8.33°$	$6.82°$	$6.00°$	$5.15°$

TABLE VII
FOUR SOURCES: THE RMSE OF THE EVALUATED METHODS IN REAL-WORLD EXPERIMENTS

True sources	TF-CHB	TF-AHB	CH-SS-PIV	DPD-MUSIC	TF-MW-BNP-AHB
$(-30°, 0°, 45°, 115°)$	$12.00°$	$12.50°$	$11.23°$	$9.75°$	$9.09°$
$(-110°, 0°\ 30°, 100°)$	$11.25°$	$9.75°$	$8.81°$	$6.75°$	$5.21°$
$(-150°, -40°, 30°, 120°)$	$8.75°$	$10.75°$	$6.85°$	$5.00°$	$3.89°$
$(-160°, -70°, 20°, 110°)$	$7.50°$	$9.00°$	$6.04°$	$4.75°$	$4.48°$

to improve parameter estimation in reverberant and noisy environments. The experimental results demonstrated that the proposed TF-MW-BNP-AHB method offers significantly better performance than the four baseline methods, in terms of the estimation accuracy of DOA and source number estimation, in a variety of acoustic conditions, including the challenging environments of high reverberation ($RT_{60} = 700$ ms) and low SNR (SNR $= 0$ dB). Furthermore, it was shown that for the speakers located at a variety of positions in the room, our proposed TF-MW-BNP-AHB offers a better consistency for the changes in the source positions as compared with the baseline methods.

APPENDIX A
DERIVATION OF THE FORMULA IN (24)-(27)

According to the theory of conjugate priors [40], the posterior probability and the prior probability belong to the same distribution family. Therefore, according to (18) and (19), we can get the joint PDF of $\hat{\theta}_i$ and $\{\mu_l, \sigma_l^2\}$

$$\begin{aligned} p(\hat{\theta}_i, \mu_l, \sigma_l^2|\boldsymbol{\Psi}_l, k_i) &= p(\hat{\theta}_i|\mu_l, \sigma_l^2, k_i)p(\mu_l, \sigma_l^2|\boldsymbol{\Psi}_l) \\ &= \frac{\gamma_l^{\eta_l} e^{-\gamma_l/\sigma_l^2}}{\Gamma(\eta_l)\sigma_l^{2(\eta_l-1)}}\left(\frac{\xi_l}{2\pi\sigma_l^2}\right)^{\frac{1}{2}} e^{\left[-\frac{\xi_l}{2\sigma_l^2}(\mu_l-\chi_l)^2\right]}\times \\ &\quad \frac{1}{\sqrt{2\pi\sigma_l^2}} e^{\frac{-(\hat{\theta}_i+2\pi k_i-\mu_l)^2}{2\sigma_l^2}} \\ &\propto \frac{1}{\sigma_l^{2\eta_l}} e^{\left[-\frac{\xi_l(\mu_l-\chi_l)^2+2\gamma_l+(\hat{\theta}_i+2\pi k_i-\mu_l)^2}{2\sigma_l^2}\right]}. \end{aligned} \tag{43}$$

According to the Bayesian rule [24], the posterior probability of the DOA estimate can be expressed as

$$p(\mu_l, \sigma_l^2|\hat{\theta}_i, \boldsymbol{\Psi}_l, k_i) \propto p(\hat{\theta}_i|\mu_l, \sigma_l^2, k_i)p(\mu_l, \sigma_l^2|\boldsymbol{\Psi}_l). \tag{44}$$

Since the mean and variance of IGMM that we established follows the Gaussian Gamma distribution, which is the conjugate prior of the Gaussian distribution [40]. According to the property of the conjugate prior [29], we know that the posterior probability also follows the Gaussian Gamma distribution and can be written as

$$p(\mu_l, \sigma_l^2|\hat{\theta}_i, \boldsymbol{\Psi}_l, k_i) \propto \boldsymbol{N}(\mu_l|\tilde{\chi}_l, \sigma_l^2/\tilde{\xi}_l)\boldsymbol{G}(\sigma_l^{-2}|\tilde{\eta}_l, \tilde{\gamma}_l), \tag{45}$$

For (45), the right side of $\propto$ can be represented as the expanded form of (19), and the left side can be simplified as the form of Gaussian Gamma distribution, therefore we can obtain

$$\begin{aligned} &\frac{1}{\sigma_l^{2\eta_l}} e^{\left\{-\frac{1}{2\sigma_l^2}\left[(\xi_l+1)\left(\mu_l-\frac{\xi_l\chi_l+\hat{\theta}_i+2\pi k_i}{\xi_l+1}\right)^2+2\left(\gamma_l+\frac{\xi_l(\hat{\theta}_i+2\pi k_i-\chi_l)^2}{2(\xi_l+1)}\right)\right]\right\}} \\ &\propto \frac{1}{\sigma_l^{2\tilde{\eta}_l-1}} e^{\left\{-\frac{1}{2\sigma_l^2}[\tilde{\xi}_l(\mu_l-\tilde{\chi}_l)^2+2\tilde{\gamma}_l]\right\}}, \end{aligned} \tag{46}$$

In (46), we know that both sides of the proportional sign follow Gaussian Gamma distribution. Thus, the update formula (24), (25), (26) and (27) can be obtained by one-to-one corresponding parameters [41].

ACKNOWLEDGEMENT

The authors wish to thank the associate editor and the anonymous reviewers for their helpful suggestions.

REFERENCES

[1] A. Marti, M. Cobos, and J. J. Lopez, "Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Prague, Czech Republic, May, 2011, pp. 2592-2595.

[2] S. Hafezi, A. H. Moore, and P. A. Naylor, "Augmented intensity vectors for direction of arrival estimation in the spherical harmonic domain" *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1956-1968, Oct. 2017.

[3] L. Sun and Q. Cheng, "Indoor sound source localization and number estimation using infinite Gaussian mixture models," in *48th Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, USA, April, 2015, pp. 1189-1193.

[4] D. S. Brungart, J. Cohen, D. J. Zion, and G. D. Romigh, "The localization of non-individualized virtual sounds by hearing impaired listeners," *J. Acoust. Soc. Am.*, vol. 141, no. 4, pp. 2870-2881, Apr. 2017.

[5] W. Kellermann, "Beamforming for speech and audio signals," in *Handbook of Signal Processing in Acoustics*, D. Havelock, S. Kuwano, and M. Vorländer, Eds., New York, NY, USA: Springer, 2008, pp. 691-702.

[6] Y. Chen, W. Wang, Z. Wang, and B. Xia, "A source counting method using acoustic vector sensor based on sparse modeling of DOA histogram," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 69-73, Jan. 2019.

[7] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," *IEEE Trans. Ind. Electron.*, vol. 65, no. 8, pp. 6403-6413, Aug. 2018.

[8] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech, Signal Mag.*, vol. 5, no. 2, pp. 4-24, Apr. 1988.

[9] S. Braun and I. Tashev, "Acoustic localization using spatial probability in noisy and reverberant environments," in *IEEE Workshop App. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2019, pp. 353-357.

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

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

[12] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 4, pp. 823-831, Aug. 1985.

[13] C. Knapp and G. 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.

[14] S. Lin, "Reverberation-robust localization of speakers using distinct speech onsets and multi-channel cross-correlations," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2098-2111, Jul. 2018.

[15] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. Eur. Signal Process. Conf.*, Aalborg, Denmark, Aug. 2010, pp. 442-446.

[16] M. J. Crocker and F. Jacobsen, "Sound intensity," in *Handbook of Acoustics*, M. J. Crocker, Ed., New York, NY, USA: Wiley-Interscience, 1998, ch. 106, pp. 1327-1340.

[17] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*, Berlin/Heidelberg, Germany: Springer-Verlag, 2007.

[18] E. Mabande, H. Sun, K. Kowalczyk, and W. Kellermann, "Comparison of subspace-based and steered beamformer-based reflection localization methods," in *Proc. 19th Eur. Signal Process. Conf.*, Barcelona, Spain, Aug. 2011, pp. 146-150.

[19] E. Tiana-Roig, F. Jacobsen, and E. F. Grande, "Beamforming with a circular microphone array for localization of environmental noise sources," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3535-3542, Dec. 2010.

[20] A. M. Torres, M. Cobos, B. Pueo and J. J. Lopez, "Robust acoustic source localization based on modal beamforming and time-frequency processing using circular microphone arrays," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1511-1520, Sep. 2012.

[21] H. Teutsch and W. Kellermann, "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2724-2736, Aug. 2006.

[22] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudo-intensity vectors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 178-192, Jan. 2017.

[23] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494-1505, Oct. 2014.

[24] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian nonparametrics for microphone array processing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 493-504, Feb. 2014.

[25] D. F. Comesana, K. R. Holland, D. G. Escribano, and H. E. D. Bree, "An introduction to virtual phased arrays for beamforming applications," *Arch. Acoust.*, vol. 39, no. 1, pp. 81-88, Feb. 2014.

[26] E. Tiana-Roig, A. Torras-Rosell, E. Fernandez-Grande, C.-H. Jeong, and F. Agerkvist, "Towards an enhanced performance of uniform circular arrays at low frequencies," in *Proc. INTER-NOISE 2013*, Innsbruck, Austria, Sep. 2013, pp. 1-10.

[27] J. D. Maynard, E. G. Williams, and Y. Lee, "Nearfield acoustic holography: I. Theory of generalized holography and the development of NAH," *J. Acoust. Soc. Am.*, vol. 78, no. 4, pp. 1395-1413, 1985.

[28] Q. Fu, M. Li, L. Wei, and D. Yang, "An improved method combining beamforming and acoustical holography for the reconstruction of the sound pressure on structure surface," *Acta Acust. United Ac.*, vol. 100, no. 1, pp. 166-183, 2014.

[29] O. Walter, L. Drude, and R. Haeb-Umbach, "Source counting in speech mixtures by nonparametric Bayesian estimation of an infinite Gaussian mixture model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 459-463.

[30] S. J. Gershman and D. M. Blei, "A tutorial on bayesian nonparametric models," *J. Maths. Psy.*, vol. 56, no. 1, pp. 1-12, Feb. 2012.

[31] 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.

[32] A. Alinaghi, P. J. Jackson, Q. Liu and W. Wang, "Joint mixing vector and binaural model based stereo source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1434-1448, Sept. 2014.

[33] X. Chen, W. Wang, Y. Wang, X. Zhong, and A. Alinaghi, "Reverberant speech separation with probabilistic time-frequency masking for B-format recordings," *Speech Comm.*, vol. 68, pp. 41-54, Apr. 2015.

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

[35] F. Jacobsen, J. Hald, E. Fernandez-Grande, and G. Moreno, "Spherical near field acoustic holography with microphones on a rigid sphere," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 3385, Jun., 2008.

[36] E. G. Williams, *Fourier Acoustics: Sound Radiation and Near Field Acoustic Holography*, London: Academic, 1999.

[37] J. Meyer and G. W. Elko, "Spherical harmonic modal beamforming for an augmented circular microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Las Vegas, NV, USA, Mar. 2008, pp. 5280-5283.

[38] A. M. Torres, J. Mateo, and M. Cobos, "Room acoustics analysis using circular arrays: A comparison between plane-wave decomposition and modal beamforming approaches," *Circuits. Syst. Signal Process.*, vol. 35, no. 5, pp. 1625-1642, May 2016.

[39] A. M. Torres, M. Cobos, B. Pueo and J. J. Lopez, "Robust acoustic source localization based on modal beamforming and time-frequency processing using circular microphone arrays," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1511-1520, Sep. 2012.

[40] X. Yu, "Gibbs Sampling Methods for Dirichlet Process Mixture Model: Technical Details," pp. 1-18, Sep. 2014. [Online]. Available: http://xiaodongyu.blogspot.com/2009/09/gibbs-sampling-for-dp-mixtures.html

[41] K. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," pp. 1-29, Oct. 2007. [Online]. Available: https://www.cs.ubc.ca/ murphyk/Papers/bayesGauss.pdf

[42] H. L. Van Trees, *Optimum Array Processing. Part IV of Detection, Estimation and Modulation Theory*, New York: Wiley, 2002.

[43] L. Sun, and Q. Cheng, "Indoor multiple sound source localization using a novel data selection scheme," in *IEEE 48th Ann. Conf. Info. Sci. Sys.*, Princeton, NJ, USA, Mar. 2014, pp. 1-6.

[44] S. He and H. Chen, Closed-form DOA estimation using first-order differential microphone arrays via joint temporal-spectral-spatial processing," *IEEE Sensors J.*, vol. 17, no. 4, pp. 1046-1060, Feb. 2017.

[45] M. Kuhne, R. Togneri, and S. Nordholm, "Robust source localization in reverberant environments based on weighted fuzzy clustering", *IEEE Signal Process. Lett.*, vol. 16, no. 2, pp. 85-88, Feb. 2009.

[46] E. A. P. Habets, "RIR Generator," 2016. [Online]. Available: https://github.com/ehabets/RIR-Generator

[47] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics", *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943-950, Apr. 1979.

[48] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM", National Institute of Standards and Technology, Gaithersburg, MD, USA, NIST Interagency/Internal Rep. 4930, Feb. 1993.

[49] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247-251, Jul. 1993.

Kunkun SongGong received the B.S. degree from Hefei Normal University, Hefei, China, in 2012, and the M.S. degree from the Nanjing University of Information Science and Technology, Nanjing, China, in 2016. Currently, he is pursuing the Ph.D. degree in communication and information systems with the Nanjing University of Aeronautics and Astronautics, Nanjing, China. Since 2021, he is a visiting Ph.D. student with the Centre for Vision Speech and Signal Processing, University of Surrey, UK.

Huawei Chen received the B.S. degree from Henan Normal University, Xinxiang, China, in 1999, and the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2002 and 2004, respectively. He is a Professor in the College of Electronic and Information Engineering, at Nanjing University of Aeronautics and Astronautics, China. He currently serves as an Associate Editor of the journal IEEE ACCESS and the Springer journal Circuits, Systems, and Signal Processing.

Wenwu Wang received the B.Sc. degree in 1997, the M.E. degree in 2000, and the Ph.D. degree in 2002, all from Harbin Engineering University, China. He is a Professor in Signal Processing and Machine Learning, at University of Surrey. He is a Senior Area Editor for IEEE Transactions on Signal Processing, an Associate Editor for IEEE/ACM Transactions on Audio Speech and Language Processing, and an Elected Member of IEEE SPTM and MLSP technical committees.