Applied Acoustics 127 (2017) 336–345

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

DOA estimation of multiple speech sources by selecting reliable local sound intensity estimates

Shaowei Ding, Huawei Chen *

College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

ARTICLE INFO

Article history:
Received 3 December 2016
Received in revised form 24 June 2017
Accepted 2 July 2017
Available online 10 July 2017

Keywords:
DOA estimation
Differential microphone array
Sound intensity

ABSTRACT

Sound source DOA estimation using first-order differential microphone arrays (DMAs) has been demonstrated as a promising means for the applications where the size of arrays is restricted. The existing methods for DOA estimation of multiple speech sources with first-order DMAs, however, are shown sensitive to noise and room reverberation. To combat the problem, we propose a DOA estimation algorithm by exploring the redundancies of two orthogonal first-order DMAs in sound intensity measurement. In particular, the reliable time–frequency points for DOA estimation can be effectively singled out by the proposed algorithm and thus leads to better DOA estimation performance in noisy and reverberant environments. Moreover, the proposed algorithm has a closed form solution, and no time-consuming search process over spatial space is required. Simulation and real experimental results have demonstrated the effectiveness of the proposed algorithm.

1. Introduction

Sound source DOA estimation using microphone arrays has found a variety of practical applications, such as automatic camera tracking for teleconferencing, beamformer steering for acoustic signal acquisition, hearing aids, and speech interface with humanoid robots, just to mention a few [1–4]. Traditional methods for sound source DOA estimation include the time-difference-of-arrival (TDOA) based, the steered response power (SRP) based, and the high-resolution spectral estimation based methods [2]. It is well known that room reverberation can significantly impair the performance of sound source DOA estimation using microphone arrays. To deal with the problem, one possible solution is to use a large microphone array with increasing number of microphones, since a large-sized array implies a higher spatial diversity. For instance, to improve the robustness of TDOA-based method, robust TDOA estimation approaches using spatial prediction [5] and more recently using joint spatio-temporal prediction [6] have been proposed, which are at a cost of increasing the number of microphones. In some practical applications, however, the number of microphones may be restricted and hence small-sized arrays are preferred.

The traditional methods for sound source DOA estimation only use sound pressure information of a sound field measured by microphone arrays. Actually, a sound field not only contains scalar information, i.e., sound pressure, but also vector information, i.e., particle velocity. By simultaneously measuring sound pressure and particle velocity, the corresponding sound intensity whose direction is linked to sound source DOA, can be estimated [7], which is useful for DOA estimation. A commonly used scheme for particle velocity measurement using microphones is the p–p measurement principle [8,9]. The key point with the p–p measurement principle is that the gradient of sound pressure can be approximated by a finite difference of sound pressures measured by two closely spaced microphones, which are constructed as a first-order differential microphone array (DMA). Since the size of a first-order DMA is small in nature, it provides a promising means for sound source DOA estimation with small-sized arrays. The sound intensity based methods using DMAs have been applied to DOA estimation of underwater acoustic sources [10,11], and more recently, to convolutive speech signal separation [12–14] where the DOAs of multiple speech signals are required, which can be estimated typically via the histogram algorithm [13]. In addition, sound intensity based DOA estimation via DMAs has also been applied to the directional audio coding (DirAC) technique, which is efficient for the analysis and reproduction of spatial sound [15]. The DMAs used in [15] is constructed by a five-element planar microphone array, with four elements around a circle and the remaining one at the center of the circle. In [16], the systematic bias analysis in DOA estimation with sound intensity based method for DirAC is conducted, and a simple yet effective approach

* Corresponding author.
E-mail address: hwchen@nuaa.edu.cn (H. Chen).

http://dx.doi.org/10.1016/j.apacoust.2017.07.002
0003-682X/

is also proposed to compensate the bias. Considering the fact that the limit of the sound intensity based DOA estimation using a square array of four omnidirectional microphones for DirAC, the study in [17] shows that reliable DOA estimation over the entire audible frequency range can be achieved through inserting a rigid cylinder into the square array. Although sound intensity is shown to be useful for DOA estimation for small arrays, it has been revealed that room reverberation may degrade the DOA estimation performance prominently [18,19].

Recent years also have seen much interest in sound intensity based DOA estimation using spherical microphone arrays [20–23]. Among the existing methods, one is originally proposed for single source scenarios [20], and the others are devised for multiple sources [21–23]. The key point to these methods lies in the construction of the so-called pseudo-intensity vector, which is inspired by the standard concept of sound intensity. The pseudo-intensity vector, pointing to the direction of sound source, is usually constructed in the spherical harmonic domain. Similar to the sound intensity based DOA estimation methods using DMAs, the pseudo-intensity vector based methods using spherical microphone arrays are also in closed-form. Therefore, in comparison to the spatial spectrum estimation based methods for spherical microphone arrays, e.g., [25–27], the pseudo-intensity vector based methods are more computationally efficient since no exhaustive search is needed in sound source DOA finding. Although related, however, we would like to point out that the DOA estimation using DMAs in the present paper is different from that using spherical microphone arrays [20–23]. Firstly, the DOA estimation with DMAs is more challenging than with spherical arrays, since the DMAs we used is a much smaller array with only 4 microphones located on a circle with a radius of 2 cm. By contrast, the spherical arrays used in [20–23] consist of 32 elements distributed on a sphere with a radius of 4.2 cm. Secondly, for our DMAs, it is a planar array and is applied for the azimuth angle estimation. While for the spherical arrays, they are 3-dimensional arrays and can be used for estimation of both the azimuth and elevation angles.

In this paper, we study the problem of sound intensity based DOA estimation for multiple speech sources using two orthogonal first-order DMAs in reverberant environment. Considering the fact that speech signals are sparse in the time-frequency (T-F) domain [28], therefore DOA estimation methods for multiple speech sources are usually performed in the T-F domain via the short-time Fourier transform (STFT). Since speech signals are sparse in the T-F domain, some T-F points may contain only background noise. Moreover, the effect of room reverberation on the received signals by microphone arrays varies over different T-F points. Consequently, only a fraction of T-F points are actually reliable and useful for the source DOA estimation. However, the existing sound intensity based DOA estimation methods for multiple speech sources [12,14] have taken all the T-F points in the whole T-F domain into consideration, which may impair their performance in noisy and reverberant environments due to the destructive impact of unreliable T-F points. To overcome this drawback, in this paper we propose a DOA estimation algorithm for multiple speech sources robust against room reverberation and additive noise. The novelty of this algorithm is that it offers an effective method on how to single out those reliable T-F points by fully exploring the redundancies of the orthogonal first-order DMAs in sound intensity measurement in order to improve DOA estimation performance in noisy and reverberant environment. The proposed method to combat room reverberation and additive noise is different from those in the pseudo-intensity vector based DOA estimation using spherical microphone arrays [20,22,23], and it is particularly tailored to the DMAs. Moreover, to the best of the authors' knowledge, no such method for DOA estimation using the DMAs is available in the literature. Another advantage of the proposed algorithm is that it has a closed form solution and thus no time-consuming search process over some spatial space is required. The effectiveness of the proposed algorithm is finally verified by the simulation and real experimental results. Recently in [29], we have proposed a DOA estimation algorithm using two orthogonal first-order DMAs via joint temporal-spectral-spatial processing. We would like to mention that, unlike the present work, the method in [29] is tailored only for the problem of *single-source* DOA estimation. Note that the beamwidth of the typical DMA spatial filters employed in [29] is usually very large, (for example, for a Cardioid response, its 3-dB beamwidth attains 131°.) Therefore, it may be no longer effective for discrimination of multiple source T-F points, since the spatial resolution is rather limited.

The rest of the paper is organized as follows. Section 2 gives a brief introduction to the signal model and fundamentals of DOA estimation via sound intensity using two orthogonal first-order DMAs. In Section 3, the proposed algorithm is presented, which consists of two stages discussed in Sections 3.1 and 3.2, respectively. The simulations results and real-world experimental results are shown in Sections 4 and 5, respectively. Finally, Section 6 concludes the paper.

2. Background

2.1. Signal model

Consider a four-element circular microphone array consisting of two orthogonal first-order DMAs, as shown in Fig. 1. The DMA along the x-axis is constructed by microphones M_1 and M_3, while the DMA along the y-axis is constructed by microphones M_2 and M_4. The size of both DMAs is D, and the geometric center of the two DMAs is chosen as the origin of the coordinate system.

Suppose that N farfield speech sources in a reverberant enclosure impinge on the array with the DOAs $\phi_1, \phi_2, \ldots, \phi_N$. Herein, the DOAs are defined with respect to the positive x-axis, which implies that $\phi_n \in [-180^\circ, 180^\circ), n = 1, \ldots, N$. The signal received at the mth microphone can be expressed as

$$p_m(t) = \sum_{n=1}^{N} h_{nm}(t) \otimes s_n(t) + n_m(t) \tag{1}$$

where $s_n(t)$ represents the nth source signal, $h_{nm}(t)$ represents the room impulse response (RIR) from the nth source to the mth microphone, $\otimes$ denotes the convolution operator, and $n_m(t)$ is the additive

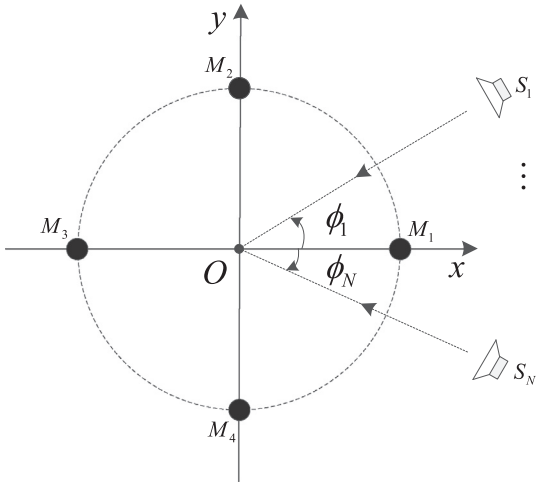


Fig. 1. Configuration of the microphone array consisting of two orthogonal first-order DMAs.

background noise that is assumed to be independent at all microphones and not correlated with the source signals.

In the STFT domain, (1) can be formulated as

$$P_m(\omega,t)=\sum_{n=1}^{N}H_{nm}(\omega,t)S_n(\omega,t)+N_m(\omega,t) \tag{2}$$

where $P_m(\omega,t), H_{nm}(\omega,t), S_n(\omega,t)$ and $N_m(\omega,t)$ stand for the STFTs of $p_m(t), h_{nm}(t), s_n(t)$ and $n_m(t)$, respectively.

It is known that speech signals are sparse in the T-F domain [28]. Thus, it leads to the common assumption that at most only one source will be active at each T-F point. Accordingly, (2) can be further reduced to

$$P_m(\omega,t)\approx H_{nm}(\omega,t)S_n(\omega,t)+N_m(\omega,t). \tag{3}$$

2.2. Instantaneous DOA estimation using sound intensity

The sound intensity based DOA estimation is based on the fact that the sound intensity is a vector quantity whose direction is linked to the sound source DOA. In the following, we briefly present the principle of the sound intensity based closed-form DOA estimation using two orthogonal first-order DMAs.

It is known that the component of particle velocity in a direction r is related to the sound pressure by [7]

$$v_r(t)=-\frac{1}{\rho}\int_{-\infty}^{t}\frac{\partial p(\tau)}{\partial r}\,\mathrm{d}\tau \tag{4}$$

where $v_r(t)$ is the particle velocity in the direction $r, p(t)$ is the sound pressure, and ρ is the density of air. In practice, the pressure gradient in (4) can be approximated by a finite difference and thus (4) becomes [7]

$$v_r(t)\approx-\frac{1}{\rho\Delta r}\int_{-\infty}^{t}[p_{r_2}(\tau)-p_{r_1}(\tau)]\mathrm{d}\tau \tag{5}$$

where p_{r_2} and p_{r_1} are sound pressures measured at two closely spaced points along the direction r, and Δr is the distance between the two points.

By applying the STFT to (5), the x- and y-components of particle velocity at the coordinate origin, measured by the first-order DMAs, can be expressed in the T-F domain, respectively, as

$$V_x(\omega,t)\approx\frac{[P_1(\omega,t)-P_3(\omega,t)]}{\omega\rho d} \tag{6}$$

$$V_y(\omega,t)\approx\frac{[P_2(\omega,t)-P_4(\omega,t)]}{\omega\rho d} \tag{7}$$

where $=\sqrt{-1}$. The sound pressure at the coordinate origin can be estimated via the average of the sound pressures at all microphones [8]

$$P_0(\omega,t)=\frac{1}{4}[P_1(\omega,t)+P_2(\omega,t)+P_3(\omega,t)+P_4(\omega,t)]. \tag{8}$$

The sound source DOA can be estimated using the active sound intensity, whose x- and y-components are given by [8,9]

$$I_x(\omega,t)=\frac{1}{2}\mathrm{Re}\{P_0(\omega,t)V_x^*(\omega,t)\} \tag{9}$$

$$I_y(\omega,t)=\frac{1}{2}\mathrm{Re}\{P_0(\omega,t)V_y^*(\omega,t)\} \tag{10}$$

where $\mathrm{Re}\{\cdot\}$ denotes the real part, and the superscript $(\cdot)^*$ denotes the complex conjugate. By (9) and (10), the instantaneous DOA at the T-F point (ω,t) can be estimated by

$$\begin{aligned}\hat{\phi}(\omega,t)&=\arctan\left[\frac{I_y(\omega,t)}{I_x(\omega,t)}\right]\\&=\arctan\left[\frac{\mathrm{Im}\{P_0(\omega,t)[P_2(\omega,t)-P_4(\omega,t)]^*\}}{\mathrm{Im}\{P_0(\omega,t)[P_1(\omega,t)-P_3(\omega,t)]^*\}}\right].\end{aligned} \tag{11}$$

3. Proposed algorithm

Our proposed algorithm consists of two stages. The first stage is to select the reliable instantaneous DOAs by local DOA variance and further by exploring the redundancies of the orthogonal DMAs to improve the robustness against room reverberation. The second stage is to estimate the DOAs of multiple sound sources in closed form via a clustering technique.

3.1. Selection of reliable instantaneous DOAs

As it is known, speech signals are sparse in the T-F domain, and some T-F points may contain only background noise [28]. Moreover, due to the presence of room reverberation, the received signals by microphones usually varies over different T-F points. As a result, only a fraction of T-F points correspond to the accurate source DOA in the estimated instantaneous DOAs $\hat{\phi}(\omega,t)$. Therefore, it has motivated us to develop the following schemes for selection of reliable instantaneous DOAs in order to improve the algorithm robustness against noise and room reverberation.

3.1.1. Preliminary selection by local DOA variance

According to [30–32], we know that the reliable instantaneous DOAs are often exhibit low local DOA variances. In contrast, high variances, however, indicate the T-F regions where room reverberations contaminate the instantaneous DOA estimation. In other words, it implies that the T-F points with low DOA fluctuations, i.e., low local DOA variance, are less affected by noise and room reverberation. Inspired by the local DOA variance, we now develop a preliminary selection scheme for reliable instantaneous DOAs using a binary masking based on the local DOA variance.

By (11), the local DOA variance at the T-F point (ω,t) can be expressed as

$$\sigma^2(\omega,t)=\frac{1}{L-1}\sum_{(\omega,t)\in\Omega_{(\omega,t)}}\left[\hat{\phi}(\omega,t)-\mu(\omega,t)\right]^2 \tag{12}$$

where $\Omega_{(\omega,t)}$ denotes the neighborhood centered about $(\omega,t), L$ is the number of T-F points within $\Omega_{(\omega,t)}$, and $\mu(\omega,t)=\frac{1}{L}\sum_{(\omega,t)\in\Omega_{(\omega,t)}}\hat{\phi}(\omega,t)$ is the local DOA mean at (ω,t). Herein, a square window of T-F points centered around (ω,t) is selected as the neighborhood $\Omega_{(\omega,t)}$. Also note that, to ensure that $\Omega_{(\omega,t)}$ is centered around (ω,t), the size of the square window, $\sqrt{L}$, should be odd.

With (12), now we can define a binary masking function as

$$B(\omega,t)=\begin{cases}1, & \text{if } \sigma^2(\omega,t)\leqslant\Gamma_1;\\ 0, & \text{otherwise.}\end{cases} \tag{13}$$

where $\Gamma_1=\eta_1\max_{\forall(\omega,t)}\{\sigma^2(\omega,t)\}, 0<\eta_1<1$, is the bound to select the reliable T-F points with low local DOA variances. Accordingly, the T-F points preliminary selected by the binary masking can be expressed as

$$\hat{\phi}'(\omega,t)=\begin{cases}\hat{\phi}(\omega,t), & \text{if } B(\omega,t)=1;\\ \text{Null}, & \text{otherwise.}\end{cases} \tag{14}$$

3.1.2. Further refinement by exploiting the redundancies of the orthogonal DMAs

In Section 2.2, the instantaneous DOAs are estimated by the orthogonal DMAs with four microphones as shown in Fig. 1. Actually, in theory, it suffices to estimate the instantaneous DOAs by just using any three microphones in Fig. 1 to construct orthogonal DMAs, i.e., there are redundancies in the four-element orthogonal DMAs. Next we show how to refine the selection of reliable instantaneous DOAs by exploiting the redundancies of the orthogonal DMAs.

3.1.2.1. Orthogonal DMAs with microphones M_2–M_1–M_4. Fig. 2 shows the three-element orthogonal DMAs consisting of microphones M_2, M_1 and M_4. Similar to (6) and (7), the x-component of particle velocity at the midpoint of microphone M_1 and M_4, and the y-component of particle velocity at the midpoint of microphone M_2 and M_1 can be estimated respectively as

$$V_x(\omega,t) \approx \frac{[P_1(\omega,t) - P_4(\omega,t)]}{\omega\rho d} \tag{15}$$

$$V_y(\omega,t) \approx \frac{[P_2(\omega,t) - P_1(\omega,t)]}{\omega\rho d}. \tag{16}$$

Since the distance from M_4 or M_2 to M_1 is small, the x- and y-components of particle velocity at microphone M_1 can be approximated respectively by (15) and (16). Therefore, the x- and y-components of the active sound intensity at microphone M_1 are given by

$$I_x(\omega,t) \approx \frac{1}{2}\mathrm{Re}\{P_1(\omega,t)V_x^*(\omega,t)\} \tag{17}$$

$$I_y(\omega,t) \approx \frac{1}{2}\mathrm{Re}\{P_1(\omega,t)V_y^*(\omega,t)\}. \tag{18}$$

By (17) and (18), the instantaneous DOAs can be estimated as

$$\begin{aligned}\hat{\phi}_1(\omega,t) &= \arctan\left[\frac{I_y(\omega,t)}{I_x(\omega,t)}\right] \\ &= \arctan\left[\frac{\mathrm{Im}\{P_1(\omega,t)[P_2(\omega,t) - P_1(\omega,t)]^*\}}{\mathrm{Im}\{P_1(\omega,t)[P_1(\omega,t) - P_4(\omega,t)]^*\}}\right].\end{aligned} \tag{19}$$

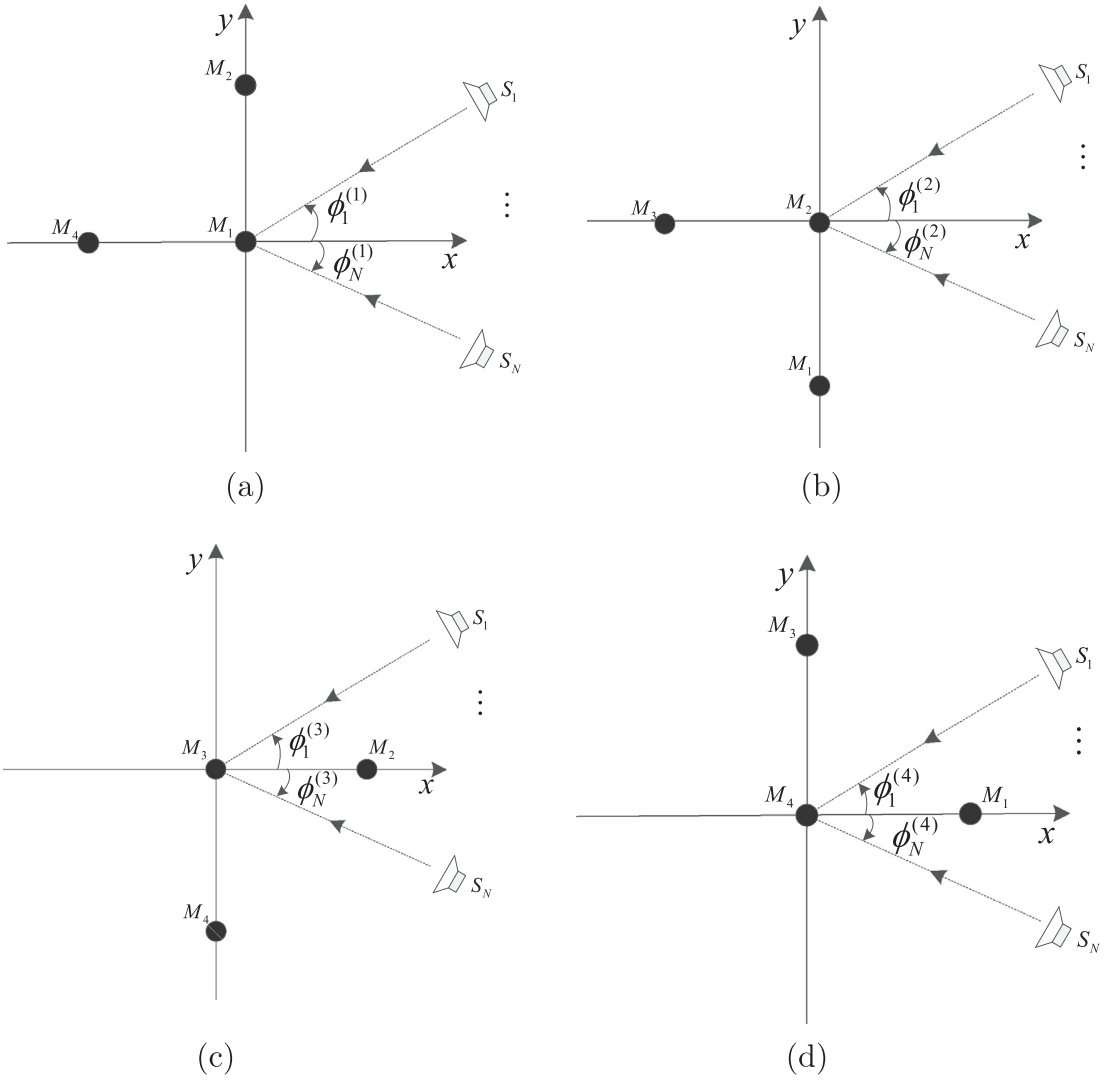


Fig. 2. Decomposition of the original four-element orthogonal first-order DMAs into four independent three-element sub-arrays, with each sub-array also constructing two orthogonal first-order DMAs which share a same microphone. (a) Orthogonal DMAs with microphones M_2–M_1–M_4. (b) Orthogonal DMAs with microphones M_3–M_2–M_1. (c) Orthogonal DMAs with microphones M_4–M_3–M_2. (d) Orthogonal DMAs with microphones M_1–M_4–M_3.

3.1.2.2. Orthogonal DMAs with microphones M_3–M_2–M_1. The three-element orthogonal DMAs consisting of microphones M_3, M_2 and M_1 are shown in Fig. 2(b). Following the similar procedures as above, by these orthogonal DMAs, the instantaneous DOAs can be estimated via

$$\begin{aligned}\hat{\phi}_2(\omega,t) &= \arctan\left[\frac{I_y(\omega,t)}{I_x(\omega,t)}\right] \\ &= \arctan\left[\frac{\mathrm{Im}\{P_2(\omega,t)[P_2(\omega,t)-P_1(\omega,t)]^*\}}{\mathrm{Im}\{P_2(\omega,t)[P_2(\omega,t)-P_3(\omega,t)]^*\}}\right].\end{aligned} \tag{20}$$

3.1.2.3. Orthogonal DMAs with microphones M_4–M_3–M_2. The three-element orthogonal DMAs consisting of microphones M_4, M_3 and M_2 are shown in Fig. 2(c). By these orthogonal DMAs, the instantaneous DOAs can be estimated via

$$\begin{aligned}\hat{\phi}_3(\omega,t) &= \arctan\left[\frac{I_y(\omega,t)}{I_x(\omega,t)}\right] \\ &= \arctan\left[\frac{\mathrm{Im}\{P_3(\omega,t)[P_4(\omega,t)-P_3(\omega,t)]^*\}}{\mathrm{Im}\{P_3(\omega,t)[P_3(\omega,t)-P_2(\omega,t)]^*\}}\right].\end{aligned} \tag{21}$$

3.1.2.4. Orthogonal DMAs with microphones M_1–M_4–M_3. The three-element orthogonal DMAs consisting of microphones M_1, M_4 and M_3 are shown in Fig. 2(d). By these orthogonal DMAs, the instantaneous DOAs can be estimated via

$$\begin{aligned}\hat{\phi}_4(\omega,t) &= \arctan\left[\frac{I_y(\omega,t)}{I_x(\omega,t)}\right] \\ &= \arctan\left[\frac{\mathrm{Im}\{P_4(\omega,t)[P_4(\omega,t)-P_3(\omega,t)]^*\}}{\mathrm{Im}\{P_4(\omega,t)[P_4(\omega,t)-P_1(\omega,t)]^*\}}\right].\end{aligned} \tag{22}$$

Notice that the coordinate systems used to derive $\hat{\phi}_i(\omega,t)$ are different from each other, and also different from the one used to derive $\hat{\phi}(\omega,t)$ by the original four-element orthogonal DMAs. Actually, $\hat{\phi}_i(\omega,t)$ can be converted to their counterparts in the coordinate system used to derive $\hat{\phi}(\omega,t)$, i.e.,

$$\hat{\phi}_i(\omega,t) \Rightarrow \begin{cases}\hat{\phi}_i(\omega,t)+45^\circ, & \text{if } \hat{\phi}_i(\omega,t)\leqslant 135^\circ;\\ \hat{\phi}_i(\omega,t)-315^\circ, & \text{otherwise.}\end{cases} \tag{23}$$

Nevertheless, this coordinate conversion is unnecessary, since we just concern the local DOA variances instead of the instantaneous DOAs, and actually there is no difference in local DOA variances with or without the coordinate conversion. Therefore, we will omit the coordinate conversion in the estimation of local DOA variances.

With $\hat{\phi}_i(\omega,t)$, the corresponding local DOA variances can be derived as

$$\sigma_i^2(\omega,t) = \frac{1}{L-1}\sum_{(\omega,t)\in\Omega_{(\omega,t)}}\left[\hat{\phi}_i(\omega,t)-\mu_i(\omega,t)\right]^2 \tag{24}$$

where $\mu_i(\omega,t)=\frac{1}{L}\sum_{(\omega,t)\in\Omega_{(\omega,t)}}\hat{\phi}_i(\omega,t)$ is the local DOA means. Similar to (13), we can define the following binary masking functions for further refining the selection of reliable T-F points as

$$B_i(\omega,t)=\begin{cases}1, & \text{if } \sigma_i^2(\omega,t)\leqslant\Gamma_2;\\ 0, & \text{otherwise.}\end{cases} \tag{25}$$

where $\Gamma_2=\eta_2\max_{\forall(\omega,t)}\{\sigma_i^2(\omega,t)\}, 0<\eta_2<1$, is the bound for refining selection of the reliable T-F points with low local DOA variances. With the binary masking functions $B_i(\omega,t),\ (i=1,2,3,4)$, we can further refine the T-F point selection by discriminating the more likely reliable T-F points from the preliminary selection in (14). A straightforward way to this end is to select those T-F points as reliable ones with all their binary masking functions equal to 1. However, our analysis has shown that this scheme may degrade the performance of DOA estimation, since the T-F point selection criterion is too stringent which may lead to many reliable T-F points being discarded. To overcome this problem, a more suitable scheme is based on the statistical point of view. If there are over one half of binary masking functions indicating a T-F point is reliable, i.e., with a probability of over 50%, then it is reasonable to consider that this T-F point is likely reliable. Mathematically, the refined selected reliable T-F points by exploiting the redundancies of the orthogonal DMAs can be expressed as

$$\hat{\phi}''(\omega,t)=\begin{cases}\hat{\phi}'(\omega,t), & \text{if } \frac{1}{4}\sum_{i=1}^{4}B_i(\omega,t)\geqslant 50\%;\\ \text{Null}, & \text{otherwise.}\end{cases} \tag{26}$$

3.2. DOA estimation via clustering

Based on the reliable T-F points $\hat{\phi}''(\omega,t)$ selected via (26), now we can estimate the DOAs of the multiple sound sources with a close-form formula. To this end, we use the celebrated k-means clustering method [34].

The cost function of the k-means clustering can be expressed as

$$J=\sum_{n=1}^{N}\sum_{\omega,t}u_n(\omega,t)\left|\hat{\phi}''(\omega,t)-c_n\right|^2 \tag{27}$$

where c_n is the cluster centroid corresponding to the nth sound source, and $u_n(\omega,t)$ is the membership coefficient which is set to 1 if a T-F point belongs to the nth sound source and to 0 otherwise.

The k-means clustering problem is solved through iteratively minimizing (27). The update equations for the membership coefficients and cluster centroids are given respectively by

$$u_n(\omega,t)=\begin{cases}1, & \text{if } |\hat{\phi}''(\omega,t)-c_n|=\min\limits_{n'=1,\ldots,N}|\hat{\phi}''(\omega,t)-c_{n'}|;\\ 0, & \text{otherwise.}\end{cases} \tag{28}$$

$$c_n=\frac{\sum_{\omega,t}u_n(\omega,t)\hat{\phi}''(\omega,t)}{\sum_{\omega,t}u_n(\omega,t)} \tag{29}$$

where the cluster centroids can be randomly initialized prior to iterations.

With the above clustering procedures, now the reliable T-F points can be classified into N clusters, denoted as C_n, which are corresponding to different sound sources. By (11), the DOA of the nth sound source can be estimated in a closed form as

$$\begin{aligned}\hat{\phi}_n(\omega,t) &= \arctan\left[\frac{\sum_{(\omega,t)\in C_n}}{I_y(\omega,t)}\sum_{(\omega,t)\in C_n}I_x(\omega,t)\right] \\ &= \arctan\left[\frac{\sum_{(\omega,t)\in C_n}\mathrm{Im}\{P_0(\omega,t)[P_2(\omega,t)-P_4(\omega,t)]^*\}}{\sum_{(\omega,t)\in C_n}\mathrm{Im}\{P_0(\omega,t)[P_1(\omega,t)-P_3(\omega,t)]^*\}}\right].\end{aligned} \tag{30}$$

Herein, to improve the robustness and also to formulate a closed-form solution, averaged active sound intensity has been used by combining the instantaneous sound intensity over various T-F points. We would like to point out that localization accuracy in low frequencies is often severely degraded by room reverberation, thus a frequency-weighting function ω_i has been added to each T-F point in deriving (30).

4. Simulation results

In this section, simulation results are presented to demonstrate the performance of the proposed algorithm by exploiting the redundancies of the orthogonal DMAs.

In the following, we will compare the proposed algorithm by exploiting the redundancies of the orthogonal DMAs with the

existing histogram algorithm for multiple sound source localization [13]. In addition, the other two related algorithms, which can be seen as the byproducts in the derivation of the proposed algorithm, are also included for comparison. One is the basic clustering algorithm which is based on the instantaneous DOAs derived in Section 2.2 without using any postprocessing, and the other is the masking-based clustering algorithm which is based on the instantaneous DOAs derived in Section 2.2, but with only the preliminary T-F points selection by local DOA variance introduced in Section 3.1.1. For ease of notation, we denote the above-mentioned four algorithms as

- **Hist**: the Histogram algorithm
- **Clust**: the basic Clustering algorithm
- **M-Clust**: the Masking-based Clustering algorithm
- **MCOR**: the proposed algorithm with Masking-based Clustering exploiting Orthogonal DMAs' Redundancies

It is noted that the single source zone detection (SSZD) has also shown effective for reliable T-F point selection for multiple speech source DOA estimation [22,24]. The main idea of the scheme is to select those T-F points where one source is dominant as the reliable T-F points. In the following, we will also compare our proposed reliable T-F point selection scheme with the SSZD based counterpart.

To facilitate algorithm evaluation, the root mean square error (RMSE), which characterizes the statistical average performance of DOA estimation, is used. The RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K}\left[\hat{\phi}_n(k) - \phi_n\right]^2} \tag{31}$$

where $\hat{\phi}_n(k)$ is the estimate of the nth source's DOA ϕ_n for the kth Monte Carlo simulation or real-world experiment, and K denotes the number of Monte Carlo simulations or real-world experiments.

4.1. Simulation setup

Consider a microphone array constructed by two orthogonal first-order DMAs as in Fig. 1, with an array size of $D = 4$ cm. The image method [33] is used to simulate a reverberant acoustic environment for a rectangular room with a size of $8 \times 6 \times 4\ \mathrm{m}^3$, and the software, RIR Generator [35], is utilized for generating the RIRs from sound source to microphones. The center of the microphone array is located at the center of the room with a height of 1.5 m above the floor. The sound source signals are selected from the well-known TIMIT speech database. All the speech source signals used are 1-s long with a sampling frequency of 16 kHz, which lie in the same plane as the array with a distance of 2 m from the center of the array. The additive noises on the microphones are mutually uncorrelated white Gaussian, and also are uncorrelated with the speech signals. Herein, $K = 100$ Monte Carlo simulations are conducted.

For all the evaluated algorithms, the STFT is calculated using a Hamming window of 512 samples length (32 ms) with 50% overlap between consecutive frames. In computing the local DOA variances the window size is chosen to be $L = 9$. Moreover, the two parameters η_1 and η_2 for reliable T-F point selection are set as 0.5 and 0.15, respectively. For the SSZD based reliable T-F point selection, the single-source zone threshold is set to $\epsilon = 0.2$, the T-F zones width is set to $K = 375$ Hz, and the frequency bins/single-source zone is set to $d = 2$ [22,24].

4.2. Effect of room reverberation

To facilitate better understanding of the proposed DOA estimation algorithm, first we present an example to demonstrate how the proposed scheme of reliable T-F point selection works in reverberant environments. Consider two speech sources, one is a female speech located at $\phi_1 = 30^\circ$ and the other is a male speech at $\phi_2 = -30^\circ$. The waveforms and spectragrams of the two speech sources are shown in Figs. 3 and 4. Here we set $T_{60} = 300$ ms and SNR = 20 dB. In this scenario, the microphone signals will be contaminated by additive noise and room reverberation. In Fig. 5 we show the signal waveform received by microphone M_1 and its spectrogram. From Fig. 5(b), we can see that the most parts of spectra of the speech sources are mainly below 6 kHz. It implies that the T-F points within 6 kHz should be singled out for DOA estimation. Fig. 6(a), (b) and (c) display the clustering results of T-F points with the Clust, M-Clust, and MCOR algorithms, respectively, where the T-F points in red color are corresponding to the female speech source, while those in blue color corresponding to the male speech source. As we can see from Fig. 6(b), the M-Clust algorithm is not sufficient in reliable T-F point selection, since so many high frequency T-F points above 6 kHz still remain. In contrast, the proposed MCOR algorithm performs well in singling out the reliable T-F points which are mainly within 6 kHz, as shown in Fig. 6(c). Comparing Fig. 6(c) with Fig. 5 we can see that the selected T-F points by the MCOR algorithm are basically consistent with those in Fig. 5 which are corresponding to the spectra of speech sources, as we desire. For the female speech source, the clustering accuracies of the Clust, M-Clust, and MCOR algorithms are 46.34%, 47.89%, and 68.30%, respectively. While for the male speech source, the clustering accuracies of the Clust, M-Clust, and MCOR algorithms are 49.81%, 50.57%, and 61.26%, respectively. Therefore, by using our proposed MCOR algorithm, the clustering accuracy can be improved and thus leads to better DOA estimation performance.

Now we analyze the effect of room reverberation on the performance of the algorithms. Consider two sound sources, with one female speech source located at $\phi_1 = 30^\circ$, and one male speech source located at $\phi_2 = -30^\circ$. In the simulations, five different sets of female and male speech signals from the TIMIT database are selected as the sound sources. For each set of speech signals, 20 Monte Carlo simulations are conducted, which results in a total number of 100 Monte Carlo simulations. Fig. 7 shows the DOA estimation RMSEs of the algorithms under various reverberation time

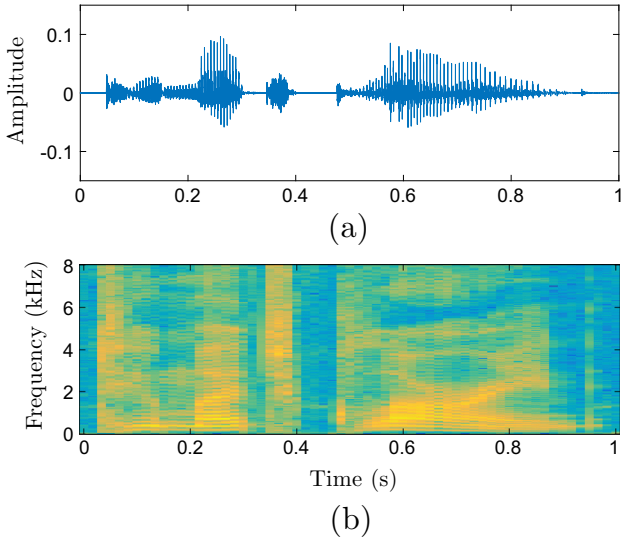


Fig. 3. The female speech signal. (a) The waveform. (b) The spectrogram.

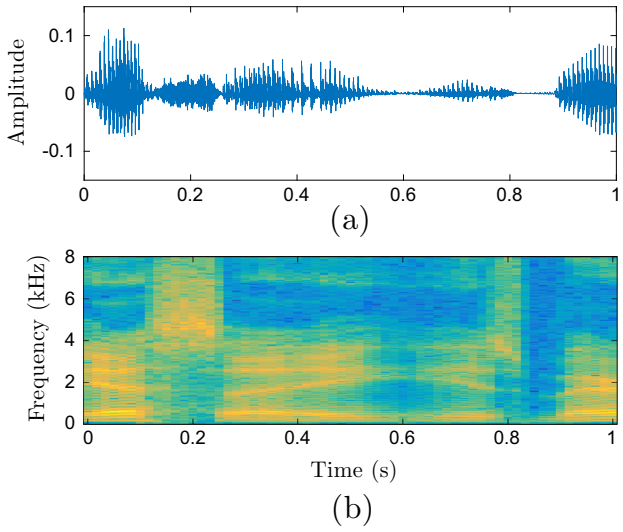


Fig. 4. The male speech signal. (a) The waveform. (b) The spectrogram.

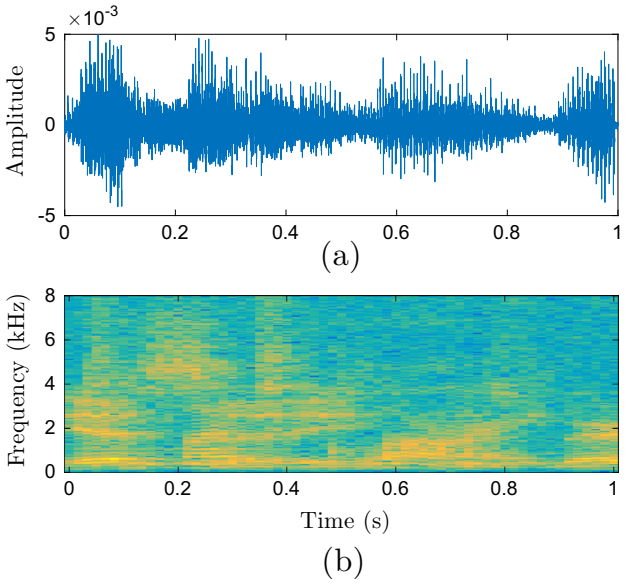


Fig. 5. The received signal of microphone M_1, where $T_{60} = 300$ ms and SNR $= 20$ dB.

T_{60}, where the signal noise ratio (SNR) is set as 15 dB. It can be seen from Fig. 7 that the basic clustering algorithm (Clust) is superior to the histogram algorithm (Hist) when the reverberation time is less than 400 ms. With the increasing of reverberation time, however, the Clust algorithm degrades worse than the Hist algorithm. In comparison, the masking-based clustering algorithm (M-Clust) exhihts a slight performance improvement over its counterpart, the Clust algorithm. It demonstrates that the preliminary reliable T-F point selection is somewhat effective to improve the robustness of the clust algorithm in reverberant environment, although the improvement is limited. The SSZD algorithm shows a slightly better performance than the M-Clust algorithm when T_{60} is less than 500 ms. However, when T_{60} is longer than 500 ms, the SSZD algorithm degrades worse than the M-Clust algorithm. This implies that the SSZD algorithm is sensitive to room reverberation. In contrast, the proposed algorithm with masking-based clustering exploiting orthogonal DMAs' redundancies (MCOR) shows the best performance among the five algorithms. Comparing with the

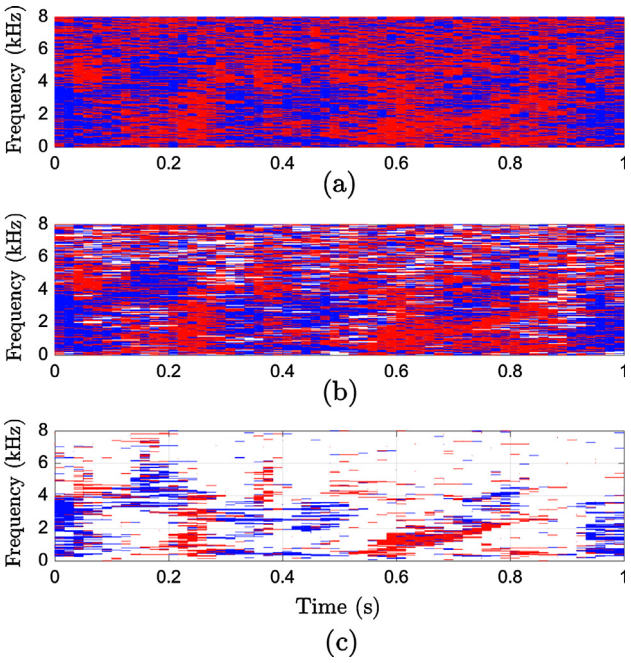


Fig. 6. The clustering results of T-F points, where T-F points in red color are corresponding to the female speech source, and those in blue color corresponding to the male speech source. (a) The Clust algorithm. (b) The M-Clust algorithm. (c) The MCOR algorithm.

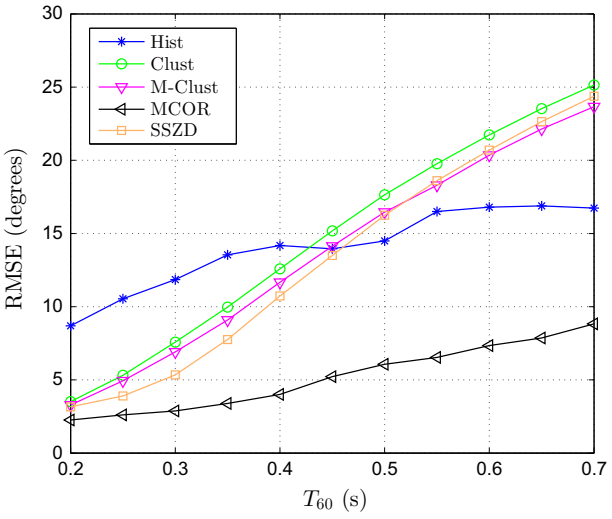


Fig. 7. DOA estimation RMSEs versus reverberation time T_{60}, where SNR $= 15$ dB.

preliminary reliable T-F point selection scheme, M-Clust, we can see that the reliable T-F point selection by exploiting the orthogonal DMAs' redundancies becomes more effective, especially in highly reverberant environment.

4.3. Effect of signal-to-noise ratio

Now we study the effect of SNR on the performance of the algorithms. Again we consider two sound sources, with one female speech source at $\phi_1 = 30^\circ$ and the other male speech source at $\phi_2 = -30^\circ$. The remaining simulation conditions are same as above, except the reverberation time is fixed at $T_{60} = 300$ ms. In Fig. 8, we plot the DOA estimation RMSEs of the algorithms as a function of SNR. Generally speaking, we can see from the

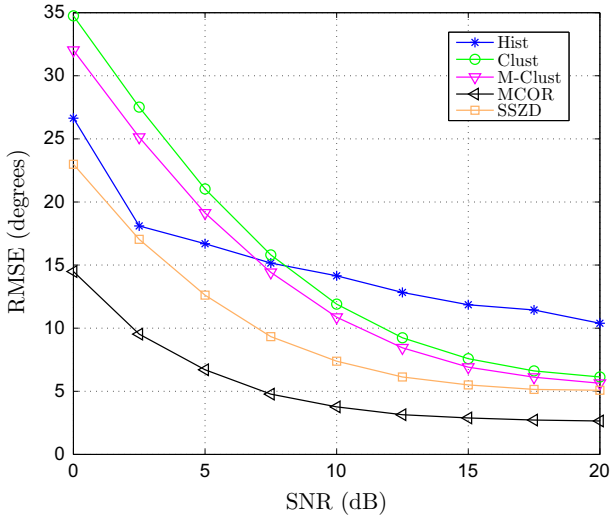


Fig. 8. DOA estimation RMSEs versus SNR, where $T_{60} = 300$ ms.

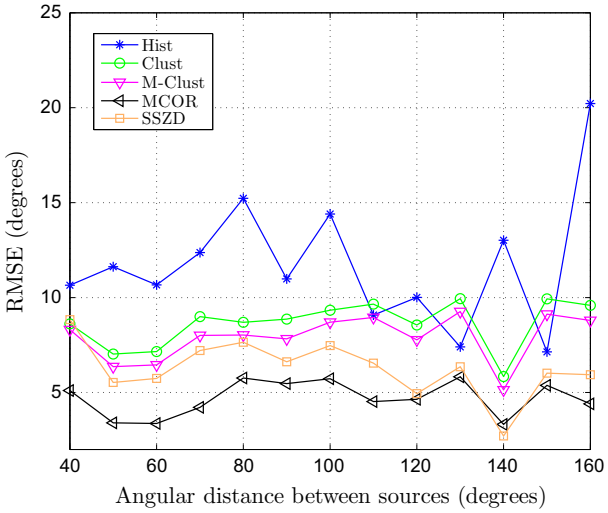


Fig. 9. DOA estimation RMSEs versus angular distance between sources, where $T_{60} = 300$ ms and SNR = 15 dB.

simulation results that the effect of SNR on the algorithms are quite similar to that of room reverberation. Explicitly, when the SNR is over 10 dB, the Clust algorithm outperforms the Hist algorithm. However, the Clust algorithm is more sensitive to the additive noise, and it will degrade worse than the Hist algorithm at low SNRs. Moreover, the M-Clust algorithm performs slightly better than the Clust algorithm, which shows that the preliminary reliable T-F point selection scheme only has limited effectiveness in combating additive noise. We can also see from Fig. 8 that the SSZD algorithm outperforms the Clust and M-Clust algorithms in the presence of additive noise, especially at low SNRs. Nevertheless, compared to the above four algorithms, our proposed MCOR algorithm is most superior and achieves the best DOA estimation performance at various SNRs. From Fig. 8 we can also see that the improvement of the MCOR algorithm over the M-Clust or Clust algorithms tends to be more significant with SNR decreasing, which implies that the reliable T-F point selection scheme exploiting the orthogonal DMAs' redundancies is highly effective to improve the algorithm robustness against additive noise.

4.4. Effect of angular distance between sources

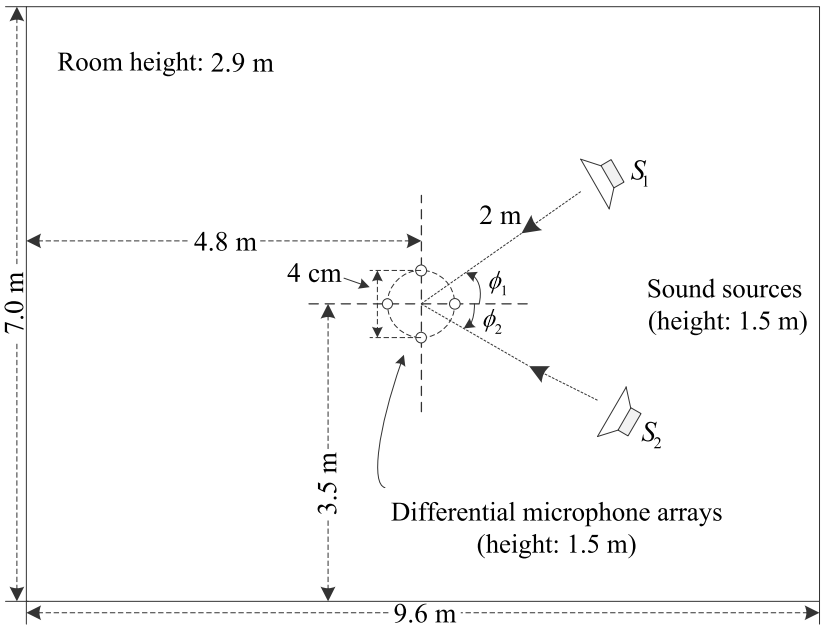
Next we study the effect of angular distance between sources on the performance of the algorithms. We consider that the male speech source fixed at $\phi_1 = -20^\circ$, while the location of the other female speech source varies from $\phi_2 = 20^\circ$ to 140° with a step size of 10°, which is corresponding to angular distances between sources from $\phi_2 = 40^\circ$ to 160°. Herein the reverberation time is set as $T_{60} = 300$ ms, and the SNR is set as 15 dB. The remaining simulation conditions are same as the above. Fig. 9 displays the DOA estimation RMSEs of various algorithms as a function of angular distance between sources. The average RMSEs over various angular distances between sources for the Hist, Clust, M-Clust, MCOR, and SSZD algorithms are $11.75^\circ, 8.63^\circ, 7.90^\circ, 4.70^\circ, 6.27^\circ$, respectively. While the maximum RMSEs over various angular distances between sources for the Hist, Clust, M-Clust, MCOR, and SSZD algorithms are $20.22^\circ, 9.95^\circ, 9.27^\circ, 5.83^\circ, 8.84^\circ$, respectively. Therefore, the proposed MCOR algorithm which exploiting the orthogonal DMAs' redundancies still performs the best among the algorithms under various angular distances between sources in terms of both average and maximum RMSEs. Moreover, the M-Clust algorithm which utilizes the preliminary reliable T-F point selection scheme consistently slightly outperforms the basic Clustering algorithm. Comparatively, the performance of the Hist algorithm is more sensitive to the angular distance between sources than its clustering based counterparts.

5. Real-world experimental results

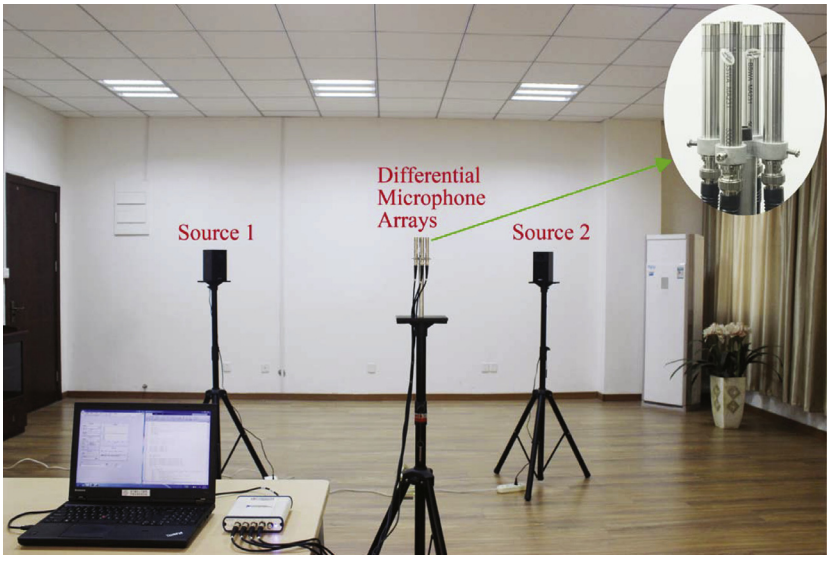
In this section, some real-world experimental results carried out in a real room are presented to illustrate the performance of the proposed algorithm.

The experiments were conducted in a rectangular room, whose size was measured as $9.6 \times 7.0 \times 2.9$ m^3. The diagram of the room and the photograph showing the experimental setup are shown in Fig. 10(a) and (b), respectively. The microphone array used in the experiments consists of two orthogonal first-order DMA with the size of 4 cm (the distance between two diagonal microphones). The microphones we used are Model MPA201 $\frac{1}{2}$-in. microphones manufactured by the BSWA Technology Co., Ltd., Beijing, China. The microphone array was placed horizontally around the center of the room with a distance of 1.5 m above the floor. The sound sources are two loudspeakers, which were also placed 1.5 m above the floor with a distance of 2 m away from the center of the microphone array. The sound source signals were recorded male English speech signals, played back simultaneously through loudspeakers. The received microphone signals were sampled with a sampling frequency of 16 kHz through a National Instruments data acquisition device (NI USB-4432) with a 24-bit resolution. For all the algorithms, the STFT is calculated using a Hamming window of 512 samples length with 50% overlap between consecutive frames. A window with the size of $L = 9$ is used in computing the local DOA variances. Moreover, the two parameters η_1 and η_2 for reliable T-F point selection are set as 0.5 and 0.15, respectively. For the SSZD algorithm, the parameter settings are same as those in the simulations above.

In the experiments, a total number of eight different sets of source DOAs were evaluated (the actual DOAs of sound sources were determined using protractors and rulers), and for each set of source DOAs, ten independent experiments have been carried out. Table 1 shows the DOA estimation RMSEs of the algorithms at various source DOAs in the real room. It can be seen from Table 1 that the experimental results are well consistent with the above simulation results. The M-Clust algorithm with the preliminary reliable T-F point selection scheme achieves slightly better performance than the basic clustering algorithm. However, when compared with the histogram algorithm, the performance of the



(a)



(b)

Fig. 10. Real-world experiments in a room. (a) Diagram of the room. (b) A photo showing the experimental setup, where a close-up view of the differential microphone arrays is also shown in the inset.

Table 1
DOA estimation RMSEs of the algorithms with real-world experiments.

(ϕ_1, ϕ_2)	Hist	Clust	M-Clust	MCOR	SSZD
$(-20°, 20°)$	10.69°	26.31°	25.55°	5.04°	28.64°
$(-20°, 30°)$	13.24°	12.37°	10.32°	2.88°	15.67°
$(-30°, 45°)$	10.22°	13.16°	7.88°	2.64°	11.15°
$(-40°, 40°)$	6.25°	8.70°	5.29°	3.86°	7.54°
$(-40°, 60°)$	8.71°	5.93°	4.47°	3.49°	9.81°
$(-50°, 75°)$	7.83°	6.19°	4.85°	4.00°	5.19°
$(-60°, 60°)$	11.21°	7.78°	5.87°	5.82°	6.02°
$(-80°, 80°)$	8.73°	6.04°	4.84°	4.30°	4.59°

M-Clust algorithm still is not satisfactory. For instance, at the case with source DOAs $(-20^\circ, 20^\circ)$, the RMSE of the M-Clust algorithm is much larger than that of the Hist algorithm, although at most cases the M-Clust algorithm has demonstrated superior performance to the Hist algorithm. We can also see from Table 1 that, the performance of the SSZD algorithm is poor in the real room environment when the DOAs of the two speech sources are close to each other, which is similar to the Clust and M-Clust algorithms. In contrast, the proposed MCOR algorithm exploiting the orthogonal DMAs' redundancies achieves the best performance for all the tested source DOAs among all the algorithms.

6. Conclusions

In this paper, we have proposed a robust DOA estimation algorithm for multiple speech sources using two orthogonal first-order DMAs. By exploring the redundancies of the two orthogonal first-order DMAs in sound intensity measurement, we have shown that the robustness of DOA estimation in noisy and reverberant environments can be effectively improved. To summarize, we have the following findings: (1) The basic clustering algorithm is superior to the existing histogram algorithm when room reverberation is low or SNR is not too low. When room reverberation is high or SNR is low, however, we arrive at the reverse conclusion. (2) By utilizing the preliminary reliable T-F point selection scheme, one can only achieve limited performance improvement in DOA estimation. (3) In contrast, by exploring the redundancies of the two orthogonal first-order DMAs in sound intensity measurement, the reliable T-F point selection becomes more effective, and hence achieves best performance among the algorithms in noisy and reverberant environments. In addition, the proposed algorithm has a closed form solution, and no time-consuming search process over spatial space is needed. The superiority of the proposed algorithm have been demonstrated through both simulations and real-world experimental results.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No. 61471190.

References

[1] Brandstein M, Ward D. Microphone Arrays: Signal Processing Techniques and Applications. Springer; 2001.
[2] Dibiase J, Silverman HF, Brandstein MS. Robust localization in reverberant rooms. In: Brstein MS, Ward DB, editors. Microphone Arrays: Signal Processing Techniques and Applications. Springer; 2001. p. 157–80.
[3] Wan X, Wu Z. Improved steered response power method for sound source localization based on principal eigenvector. Appl. Acoust. 2010;71 (December):1126–31.
[4] Lee S, Park Y, Choi J-S. Estimation of multiple sound source directions using artificial robot ears. Appl. Acoust. 2014;77(March):49–58.
[5] Chen J, Benesty J, Huang Y. Robust time delay estimation exploiting redundancy among multiple microphones. IEEE Trans. Speech Audio Process 2003;11(6):549–57.
[6] He H, Wu L, Lu J, Qiu X, Chen J. Time difference of arrival estimation exploiting multichannel spatio-temporal prediction. IEEE Trans. Audio Speech Lang. Process 2013;21(3):463–75.
[7] Fahy FJ. Sound Intensity. second ed. London, U.K.: E & FN Spon; 1995.
[8] Jacobsen F. Sound intensity. In: Rossing TD, editor. Springer Handbook of Acoustics. Springer; 2014. p. 1093–114.
[9] Gade S. Sound intensity, Part I: Theory. Brüel Kjær Tech. Rev. 1982;3:3–39.
[10] Hickling R, Wei W. Finding the direction of a sound source using a vector sound-intensity probe. J. Acoust. Soc. Am. 1993;94(4):2408–12.
[11] Hickling R, Wei W. Use of pitch-azimuth plots in determining the direction of a noise source in water with a vector sound-intensity probe. J. Acoust. Soc. Am. 1995;97(2):856–66.
[12] Gunel B, Hacihabiboglu H, Kondoz AM. Acoustic source separation of convolutive mixtures based on intensity vector statistics. IEEE Trans. Audio Speech Lang. Process 2008;16(4):748–56.
[13] X. Zhong, X. Chen, W. Wang, A. Alinaghi, A.B. Premkumar, Acoustic vector sensor based reverberant speech separation with probabilistic time-frequency masking, in: Proc. 21st European Signal Processing Conference (EUSIPCO), September 2013.
[14] Chen X, Wang W, Wang Y, Zhong X, Alinaghi A. Reverberant speech separation with probabilistic time-frequency masking for B-format recordings. Speech Commun. 2015;68(April):41–54.
[15] F. Kuech, M. Kallinger, R. Schultz-Amling, G. Del Galdo, J. Ahonen, V. Pulkki, Directional audio coding using planar microphone arrays, in: Proc. 2008 Hands-Free Speech Communication and Microphone Arrays (HSCMA), 2008, pp. 37–40.
[16] M. Kallinger, F. Kuech, R. Schultz-Amling, G. Del Galdo, J. Ahonen, V. Pulkki, Enhanced direction estimation using microphone arrays for directional audio coding, in: Proc. 2008 Hands-Free Speech Communication and Microphone Arrays (HSCMA), 2008, pp. 45–48.
[17] Ahonen J, Del Galdo G, Kuech F, Pulkki V. Directional analysis with microphone array mounted on rigid cylinder for directional audio coding. J. Audio Eng. Soc. 2012;60(5):311–24.
[18] Levin D, Habets EAP, Gannot S. On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields. J. Acoust. Soc. Am. 2010;128(4):1800–11.
[19] D. Levin, E.A.P. Habets, S. Gannot, Impact of source signal coloration on intensity vector based DOA estimation, in: Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC), August 2010.
[20] D.P. Jarrett, E.A.P. Habets, P.A. Naylor, 3D source localization in the spherical harmonic domain using a pseudointensity vector, in: Proc. 18th European Signal Processing Conference (EUSIPCO), August 2010.
[21] C. Evers, A.H. Moore, P.A. Naylor, Multiple source localisation in the spherical harmonic domain, in: Proc. 14th International Workshop on Acoustic Signal Enhancement (IWAENC), September 2014.
[22] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, A. Mouchtaris. 3D localization of multiple sound sources with intensity vector estimates in single source zones, in: Proc. 23rd European Signal Processing Conference (EUSIPCO), August 2015.
[23] A.H. Moore, C. Evers, P.A. Naylor, D.L. Alon, B. Rafaely. Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test, in: Proc. 23rd European Signal Processing Conference (EUSIPCO), August 2015.
[24] Pavlidi D, Griffin A, Puigt M, Mouchtaris A. Real-time multiple sound source localization and counting using a circular microphone array. IEEE/ACM Trans. Audio Speech Lang. Process 2013;21(August):2193–206.
[25] Li X, Yan S, Ma X, Hou C. Spherical harmonics MUSIC versus conventional MUSIC. Appl. Acoust. 2011;72(September):646–52.
[26] Nadiri O, Rafaely B. 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 2014;22(October):1494–505.
[27] S. Delikaris-Manias, D. Pavlidi, V. Pulkki, A. Mouchtaris, 3D localization of multiple audio sources utilizing 2D DOA histograms, in: Proc. 24rd European Signal Processing Conference (EUSIPCO), August 2016.
[28] Yilmaz Ö, Rickard S. Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process 2004;52(7):1830–47.
[29] He S, Chen H. Closed-form DOA estimation using first-order differential microphone arrays via joint temporal-spectral-spatial processing. IEEE Sens. J. 2017;17(4):1046–60.
[30] Abrard F, Deville Y. A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. Sign. Process. 2005;85:1389–403.
[31] Kühne M, Togneri R, Nordholm S. A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation. Sign. Process. 2010;90:653–69.
[32] Kühne M, Togneri R, Nordholm S. Robust source localization in reverberant environments based on weighted fuzzy clustering. IEEE Sign. Process. Lett. 2009;16(2):85–8.
[33] Allen JB, Berkley DA. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 1979;65(April):943–50.
[34] Theodoridis S, Koutroumbas K. Pattern Recognition. fourth ed. Academic Press; 2009.
[35] [Online]. Available: <https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator>.