Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays

Nian Liu, Huawei Chen, Kunkun Songgong, et al.

Citation: The Journal of the Acoustical Society of America **149**, 1069 (2021); doi: 10.1121/10.0003445
View online: https://doi.org/10.1121/10.0003445
View Table of Contents: https://asa.scitation.org/toc/jas/149/2
Published by the Acoustical Society of America

JASA ARTICLE

Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays[a)]

Nian Liu, Huawei Chen,[b)] Kunkun Songgong, and Yanwen Li
College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ABSTRACT:
Sound source localization in noisy and reverberant rooms using microphone arrays remains a challenging task, especially for small-sized arrays. Recent years have seen promising advances on deep learning assisted approaches by reformulating the sound localization problem as a classification one. A key to the deep learning-based approaches lies in extracting sound location features effectively in noisy and reverberant conditions. The popularly adopted features are based on the well-established generalized cross correlation phase transform (GCC-PHAT), which is known to be helpful in combating room reverberation. However, the GCC-PHAT features may not be applicable to small-sized arrays. This paper proposes a deep learning assisted sound localization method using a small-sized microphone array constructed by two orthogonal first-order differential microphone arrays. An improved feature extraction scheme based on sound intensity estimation is also proposed by decoupling the correlation between sound pressure and particle velocity components in the whitening weighting construction to enhance the robustness of the time-frequency bin-wise sound intensity features. Simulation and real-world experimental results show that the proposed deep learning assisted approach can achieve higher spatial resolution and is superior to its state-of-the-art counterparts using the GCC-PHAT or sound intensity features for small-sized arrays in noisy and reverberant environments. https://doi.org/10.1121/10.0003445

(Received 7 November 2020; revised 14 January 2021; accepted 14 January 2021; published online 10 February 2021)
[Editor: Peter Gerstoft] Pages: 1069–1084

I. INTRODUCTION

Sound source localization using multiple microphone signals is an active research topic in the microphone array processing field and has found a wide range of practical applications, such as automatic camera tracking for teleconferencing,[1] human-robot interaction,[2] and hearing aids,[3] among many others.[4–6] The well-known classical methods for sound source localization include (1) the time-difference-of-arrival (TDOA)-based,[7,8] (2) the steered-response power (SRP)-based,[9,10] and (3) the high-resolution spectral estimation-based.[11,12] These classical methods, however, may face some challenges when used in enclosed environments. One of the challenges is room reverberation due to multiple sound reflections from room boundaries and objects. These sound reflections lead to multiple attenuated and delayed replicas of the source signal in the observation signals by microphones and, hence, may introduce severe signal distortions.[13] It is known that the spatial resolution of a microphone array usually depends on its aperture size, i.e., the larger the aperture size, the higher the spatial resolution.[13] With a higher spatial resolution, we can usually also achieve more noise reduction. Unfortunately, the aperture size of a microphone array may need to be restricted in some applications. Therefore, the limited aperture size of a microphone array poses another challenge to the classical sound localization methods.

To address the challenge brought by room reverberation, machine learning-based methods have been applied to sound source localization using microphone arrays.[14–27] Traditional microphone array-based approaches, such as the aforementioned TDOA- and SRP-based methods, usually only depend on the spatial information embedded in received sensor signals to estimate sound source location. In contrast, the machine learning-based methods can also take into account the prior information related to the sound source location during training process, such as the acoustic environments, so as to better solve the sound source localization problem in the presence of high room reverberation. Nevertheless, many traditional machine learning approaches are typically suitable for the case when limited training data are available, because they suffer from high computational cost when the number of training examples is large.[28] Therefore, the sound source localization approaches based on deep learning, a form of machine learning that enables computers to learn from experience,[28] have particularly gained much interest in recent years.

As we know, the generalized cross correlation phase transform (GCC-PHAT) is robust against room reverberation to some extent, and thus it has been widely combined with the machine learning-based approaches for sound source localization.[16–23] For example, Xiao *et al.*[16] formulated the task of sound localization as a classification problem to estimate sound positions in noisy and reverberant

[a)]This paper is part of a special issue on Machine Learning in Acoustics.
[b)]Also at: State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China. Electronic mail: hwchen@nuaa.edu.cn

environments, where a multi-layer perceptron (MLP) was employed as the network architecture and the GCC-PHAT features as inputs to the MLP. Subsequently, convolutional neural networks (CNNs) have been used widely in sound source localization and have been shown to be able to produce state-of-the-art performance in classification tasks because they are able to identify a large number of local features in the inputs through shared weights over small local receptive fields.[28] Among them, Yue *et al.*[17] proposed to use the GCC-PHAT features as the inputs to a CNN to estimate the three-dimensional sound locations. In Ref. 18, Li *et al.* proposed a method by combining CNN and long short term memory (LSTM) to address online sound localization, in which the feature matrices based on the GCC-PHAT over all sensor pairs are summed up. This method is shown to be robust to the topology of microphone arrays. We would like to point out that although the deep learning-based sound localization methods using the GCC-PHAT features have been reported to be able to achieve favorable performance under reverberant environments, unfortunately, as we will show later in this paper, the GCC-PHAT features actually may not be applicable to small-sized microphone arrays.

In this paper, we are interested in sound source localization in noisy and reverberant environments with a small-sized microphone array. Toward this end, we propose a CNN assisted sound localization method using a small-sized microphone array constructed by two orthogonal first-order differential microphone arrays (DMAs). This work can be seen as an improvement over our previous work,[25] where a sound intensity (SI) estimation-based feature extraction approach was proposed with the least-squares support vector machine (LSSVM) being employed as the machine learning model. Compared with the previous work, the main contributions of the current work are briefly summarized as follows:

- One problem with the previous work is that the SI features are averaged over all the time-frequency (T-F) bins to reduce the data dimension to facilitate postprocessing by the LSSVM model. However, this leads to the complete loss of useful local information on source location over the T-F domain. In contrast, in this paper, we utilize a CNN in this work for sound source localization, which can handle high-dimensional data and is great for capturing local information, and thus can retain all the T-F bin-wise SI features over the whole T-F domain.
- As analyzed in this paper, another problem with the previous work is that the phase transform (PHAT)-based whitening weighting used in the SI feature extraction to address room reverberation is sensitive to additive noise, due to the correlation between the sound pressure and particle velocity components in the whitening weighting. To deal with the problem, an improved whitening weighting for SI features is proposed in this paper by decoupling the correlation between the sound pressure and particle velocity components in the whitening weighting construction.
- DMAs, as used in the SI feature extraction in the paper, are known to be sensitive to sensor mismatches, such as microphone gain and phase errors. However, the effect of sensor mismatches on the sound localization approaches using SI features is not clear yet. In this paper, the effect of sensor mismatches on various SI features-based sound localization approaches is studied. It is shown that the proposed sound localization method is robust against sensor mismatches, while the existing counterpart may fail to work.

Extensive simulations under various acoustic conditions as well as real-world experiments have been conducted, which unanimously show that the proposed deep learning assisted approach can achieve higher spatial resolution and is superior to its state-of-the-art counterparts using the GCC-PHAT or SI features when applied for sound localization with small-sized arrays in noisy and reverberant environments.

The rest of the paper is organized as follows. Section II gives a brief introduction to the signal model and the fundamental for SI estimation. Section III presents the framework of the proposed sound source localization system, where the feature extraction and network structure are discussed in detail in Secs. III A and III B, respectively. The simulation and real experimental results are shown in Secs. IV and V, respectively, to compare our proposed method with its baseline counterparts. Finally, Sec. VI concludes the paper.

II. SIGNAL MODEL

Consider a four-element microphone array consisting of two orthogonal first-order DMAs, as shown in Fig. 1. The omnidirectional microphones M_1 and M_3 form a first-order DMA along the x axis, and another pair of omnidirectional microphones M_2 and M_4 form a first-order DMA along the y axis. The size of both DMAs is denoted as d, and the center of the two DMAs is chosen as the coordinate origin. Suppose that there is a far-field sound source in a reverberant room impinging on the microphone array with the direction of arrival (DOA) $\theta \in [-180^\circ\ 180^\circ)$, where the DOA is defined with respect to the positive x axis. Then the signal received at the ith microphone can be expressed as

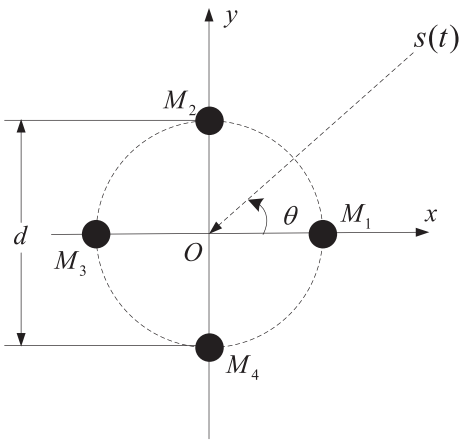

FIG. 1. Configuration of the microphone array constructed by two orthogonal first-order DMAs.

https://doi.org/10.1121/10.0003445

$$p_i(t) = h_i(t) * s(t) + n_i(t), \tag{1}$$

where $s(t)$, $h_i(t)$, and $n_i(t)$ denote the source signal, the room impulse response (RIR) from the sound source to the ith microphone, and the additive noise, respectively, and $*$ refers to the convolution operator.

The sound pressure at the coordinate origin can be estimated by averaging the received signals from all microphones,[29] i.e.,

$$p_0(t) = [p_1(t) + p_2(t) + p_3(t) + p_4(t)]/4. \tag{2}$$

By the conservation of momentum (cf. Eq. (3.12) in Ref. 30), it follows that the component of particle velocity in a direction r is related to the sound pressure by

$$v_r(t) = -\frac{1}{\rho}\int_{-\infty}^{t} \frac{\partial p(\tau)}{\partial r} d\tau, \tag{3}$$

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 Eq. (3) can be approximated by a finite difference,[29] and thus Eq. (3) becomes

$$v_r(t) \approx -\frac{1}{\rho \Delta r}\int_{-\infty}^{t} [p_{r_2}(\tau) - p_{r_1}(\tau)] d\tau, \tag{4}$$

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

By applying short-time Fourier transform (STFT) to Eq. (4), the two orthogonal velocity components at the coordinate origin, measured by the two first-order DMAs, can be expressed in the T-F domain, respectively, as

$$V_x(\omega, t) = \frac{j[P_3(\omega, t) - P_1(\omega, t)]}{\omega \rho d}, \tag{5}$$

$$V_y(\omega, t) = \frac{j[P_4(\omega, t) - P_2(\omega, t)]}{\omega \rho d}, \tag{6}$$

where $j = \sqrt{-1}$ is the imaginary unit, and

$$P_i(\omega, t) = H_i(\omega, t) S(\omega, t) + N_i(\omega, t), \tag{7}$$

where $P_i(\omega, t)$, $H_i(\omega, t)$, $S(\omega, t)$, and $N_i(\omega, t)$ stand for the STFTs of $p_i(t)$, $h_i(t)$, $s(t)$, and $n_i(t)$, respectively.

The STFT of the sound pressure at the origin, i.e., $p_0(t)$ in Eq. (2), is given by

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

With Eqs. (5), (6), and (8), the x and y components of instantaneous complex SI can be expressed as[31]

$$I_{0x}(\omega, t) = P_0(\omega, t) V_x^{\star}(\omega, t), \tag{9}$$

$$I_{0y}(\omega, t) = P_0(\omega, t) V_y^{\star}(\omega, t), \tag{10}$$

where the superscript asterisk represents the complex conjugate.

III. FEATURE EXTRACTION AND NETWORK ARCHITECTURE

In this work, the sound source localization is formulated as a classification problem by using a CNN. The diagram of the proposed sound source localization system is shown in Fig. 2. It includes two main components:

- *Feature extraction:* To solve the problem of sound source localization with the small-sized array, we propose a feature extraction method based on the SI estimation, whose direction is related to the DOA of the sound source. To improve the robustness against room reverberation and additive noise, our proposed features utilize a noise-robust whitening weighting scheme and also incorporate the redundancies in SI estimation via the four subarrays formulated through the decomposition of the original two orthogonal first-order DMAs. The extracted SI features across all T-F bins then form the feature matrices with a fixed size, which are taken as the inputs to feed the CNN. The flow chart of the feature extraction is shown in Fig. 2(b), whereas the subarray configurations are presented in detail in Fig. 3 below.
- *Network architecture:* We employ a CNN as the machine learning model. The CNN architecture used in this work is shown in Fig. 2(c). The CNN is designed to learn the mapping relationship from the proposed SI features extracted from the microphone array signals to the sound source DOA using a large set of labeled training data. Accordingly, the sound source localization problem is transformed into an end-to-end multi-classification problem.

In the following, we introduce the above two components in more detail.

A. Proposed feature extraction

1. Feature extraction based on the full-scale array

According to the theory on SI,[31] only the real part of the complex SI, i.e., the active intensity vector, contains the location information of a sound source. Therefore, the x and y components of the active SI, i.e., corresponding to the real parts of Eqs. (9) and (10), respectively, can serve as source location features. Accordingly, the extracted SI features can be expressed as

$$I_x(\omega, t) = Re\{P_0(\omega, t) V_x^{\star}(\omega, t)\}, \tag{11}$$

$$I_y(\omega, t) = Re\{P_0(\omega, t) V_y^{\star}(\omega, t)\}, \tag{12}$$

where $Re\{\cdot\}$ represents the real part.

The SI features are known to be sensitive to room reverberation and hence will lead to poor performance of sound source localization in reverberant environments. To improve the robustness against room reverberation, the existing feature extraction scheme based on the first-order DMAs[25] is proposed to apply a whitening weighting, i.e., the PHAT weighting, to the SI features. However, this is at the cost of increasing the sensitivity to additive noise. Here, we give an

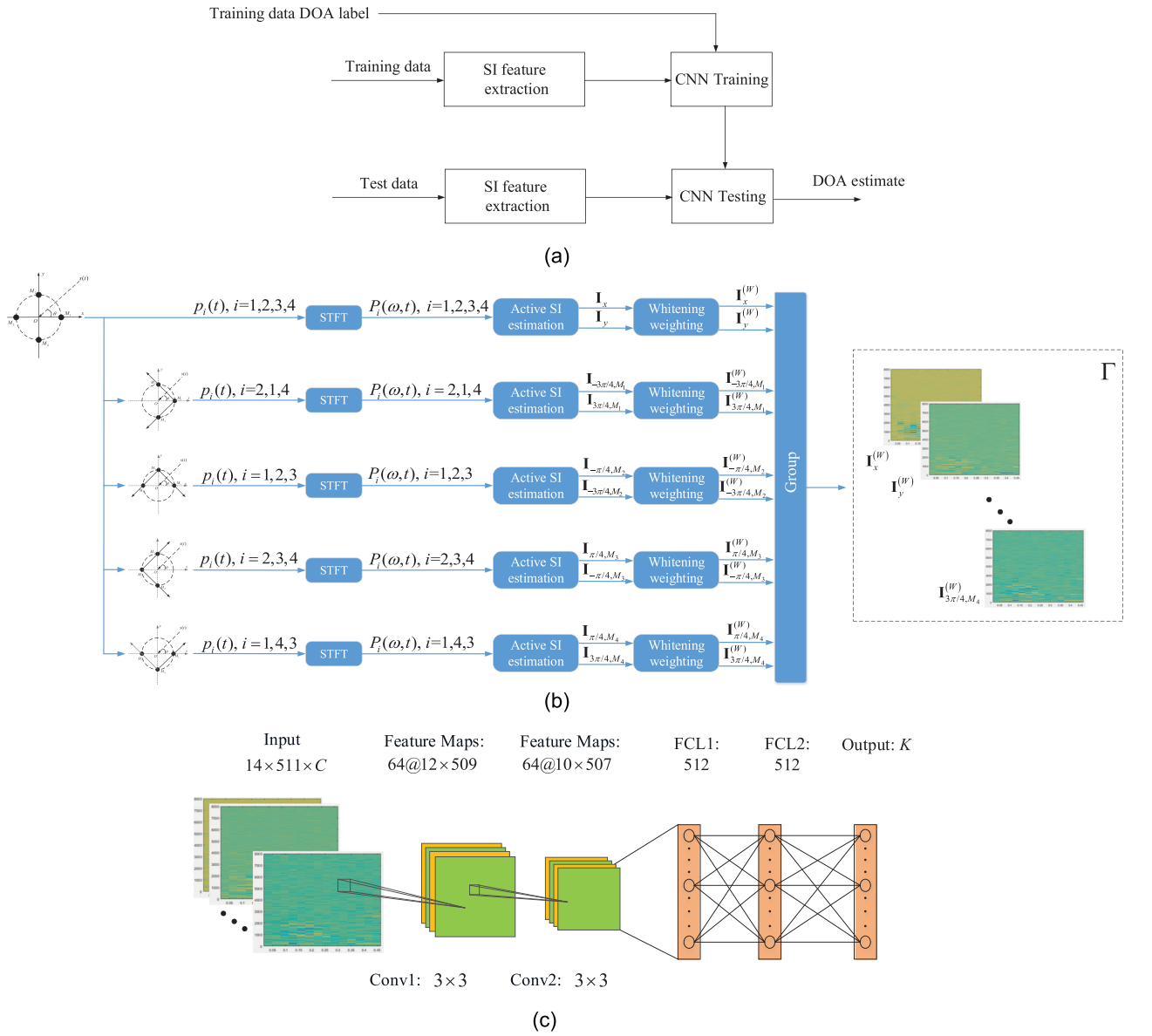

FIG. 2. (Color online) Diagram of the proposed sound source localization system: (a) overall framework; (b) feature extraction flow chart, where the subarray configurations can be seen in detail in Fig. 3 below; and (c) network architecture, where C denotes the number of channels of input features and K refers to the total number of classes.

analysis to justify this point. According to Ref. 25, the existing whitening weightings are defined as

$$W_x(\omega,t)=\frac{1}{|P_0(\omega,t)V_x^{\star}(\omega,t)|}, \tag{13}$$

$$W_y(\omega,t)=\frac{1}{|P_0(\omega,t)V_y^{\star}(\omega,t)|}, \tag{14}$$

where $W_x(\omega,t)$ and $W_y(\omega,t)$ denote the weighting for Eqs. (11) and (12), respectively. Recall from Sec. II that the sound pressure signal $P_0(\omega,t)$ is estimated via averaging over all the four microphone signals, and the two orthogonal particle velocity components $V_x(\omega,t)$ and $V_y(\omega,t)$ are estimated by the difference between the microphone signals $p_1(t)$ and $p_3(t)$ and by the difference between $p_2(t)$ and $p_4(t)$, respectively. From Eqs. (13) and (14), we can see that the whitening weightings in the existing feature extraction are formed based on the correlation between the sound pressure signal and the particle velocity components. As a result, the whitening weighting $W_x(\omega,t)$ will be highly affected by the additive noise in $p_1(t)$ and $p_3(t)$, while $W_y(\omega,t)$ will be highly affected by the additive noise in $p_2(t)$ and $p_4(t)$ under low signal-to-noise ratio (SNR) conditions.

To overcome the effect of additive noise, we have to decouple the correlation between the sound pressure and the particle velocity components in the construction of the whitening weighting. To this end, we propose to use the following whitening weighting,

$$W(\omega,t)=\left[|P_0(\omega,t)|^2+\beta(|V_x(\omega,t)|^2+|V_y(\omega,t)|^2)\right]^{1/2}, \tag{15}$$

https://doi.org/10.1121/10.0003445

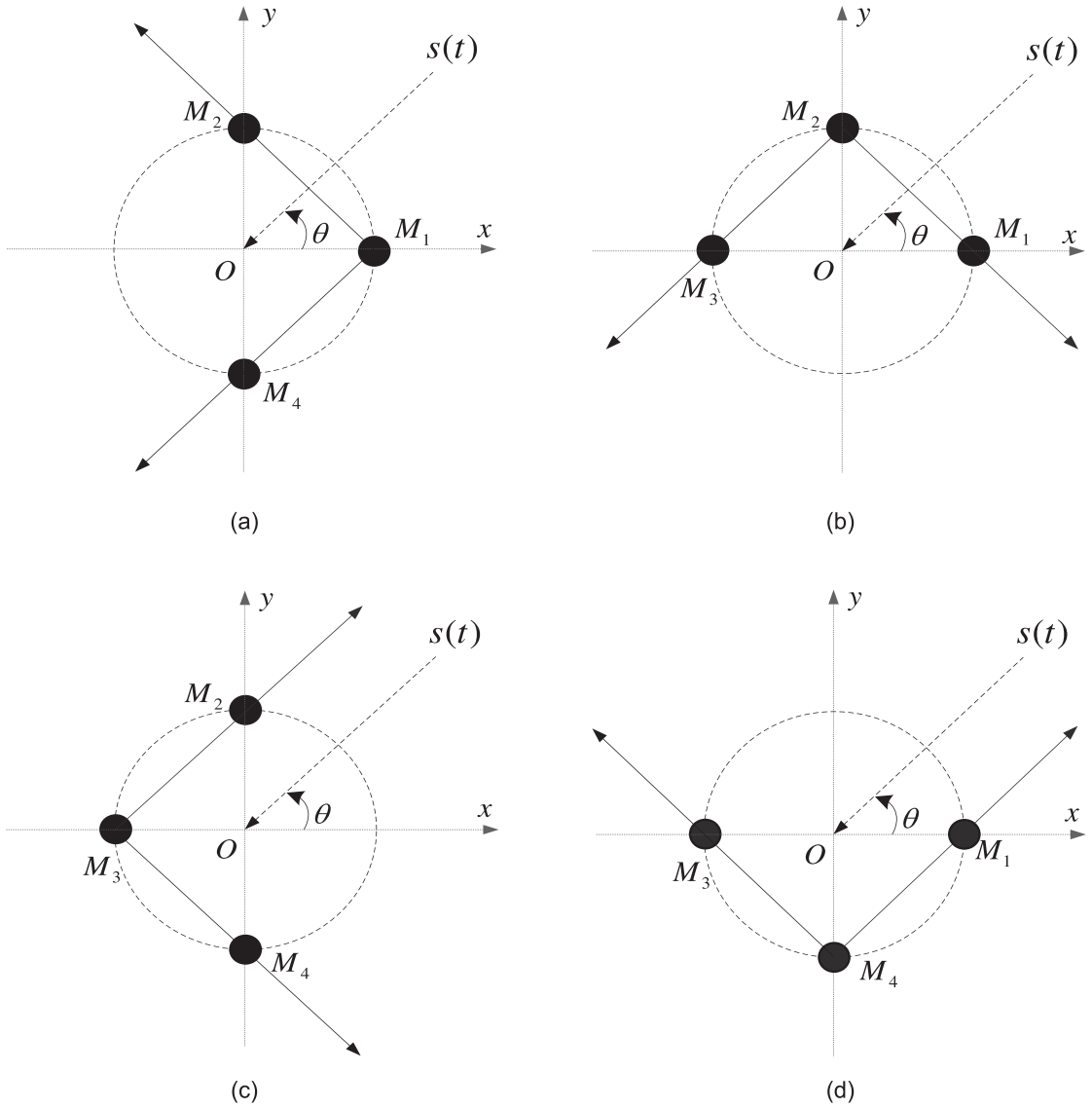

FIG. 3. Configuration of the decomposed four subarrays: (a) subarray M_2-M_1-M_4, (b) subarray M_3-M_2-M_1, (c) subarray M_4-M_3-M_2, and (d) subarray M_1-M_4-M_3.

where $\beta > 0$ denotes a trade-off parameter to be tuned in the training stage. Unlike the existing whitening weighting, the correlation between the sound pressure and the particle velocity components has now been decoupled in Eq. (15). It is noted that the power spectrum of the particle velocity components estimated via first-order DMAs is usually far less than that of the estimated sound pressure. For example, as shown in the simulation results below, their difference may be up to 6 orders of magnitude. Therefore, for our problem at hand, the setting of β should not be small. Otherwise, the two orthogonal particle velocity components may be almost ignored in the weighting.

To better understand the effect of additive noise on the existing whitening weightings, i.e., Eqs. (13) and (14), and the proposed whitening weighting, i.e., Eq. (15), now we perform some analytical analysis. To simplify notations, herein we temporally omit the arguments, (ω, t). First, we analyze the effect of additive noise on the existing whitening. By Eqs. (5), (6), and (8), Eqs. (13) and (14) can be expressed as

$$W_x = 4\omega\rho d\left\{\left|(\Sigma_{i=1}^4 H_i)^2 (H_3 - H_1)^2 S^4 + \Delta_1\right|\right\}^{-1/2}, \quad (16)$$

$$W_y = 4\omega\rho d\left\{\left|(\Sigma_{i=1}^4 H_i)^2 (H_4 - H_2)^2 S^4 + \Delta_2\right|\right\}^{-1/2}, \quad (17)$$

where Δ_1 and Δ_2 are the disturbance terms due to the presence of additive noise, which are given by

$$\Delta_1 = N_1^4 + N_3^4 + O_1(N_i), \quad (18)$$

$$\Delta_2 = N_2^4 + N_4^4 + O_2(N_i), \quad (19)$$

with $O_1(N_i)$ and $O_2(N_i)$ denoting the lower-order remainder terms. Omitting the lower-order remainder terms, the effect of the additive noise on the existing whitening is approximately characterized by

$$\frac{\Delta_1}{(\Sigma_{i=1}^4 H_i)^2(H_3-H_1)^2S^4} \simeq \frac{\left(\frac{N_1}{S}\right)^4+\left(\frac{N_3}{S}\right)^4}{(\Sigma_{i=1}^4 H_i)^2(H_3-H_1)^2}, \quad (20)$$

$$\frac{\Delta_2}{(\Sigma_{i=1}^4 H_i)^2(H_4-H_2)^2S^4} \simeq \frac{\left(\frac{N_2}{S}\right)^4+\left(\frac{N_4}{S}\right)^4}{(\Sigma_{i=1}^4 H_i)^2(H_4-H_2)^2}, \quad (21)$$

which implies the impact of the additive noise on the existing whitening depends inversely on the fourth power of SNR.

Next, we analyze the effect of additive noise on the proposed whitening. By Eqs. (5), (6), and (8), Eq. (15) can be reformulated as

$$W = \left\{\frac{1}{16}|(\Sigma_{i=1}^4 H_i)^2S^2+\Delta\varpi_1| + \frac{\beta}{\rho^2\omega^2d^2}(|(H_3-H_1)^2S^2+\Delta\varpi_2| + |(H_4-H_2)^2S^2+\Delta\varpi_3|)\right\}^{1/2}, \quad (22)$$

where the disturbance terms $\Delta\varpi_1$, $\Delta\varpi_2$, and $\Delta\varpi_3$ due to the additive noise are given by

$$\Delta\varpi_1 = N_1^2+N_2^2+N_3^2+N_4^2+O_{\varpi_1}(N_i), \quad (23)$$

$$\Delta\varpi_2 = N_1^2+N_3^2+O_{\varpi_2}(N_i), \quad (24)$$

$$\Delta\varpi_3 = N_2^2+N_4^2+O_{\varpi_3}(N_i), \quad (25)$$

with $O_{\varpi_1}(N_i)$, $O_{\varpi_2}(N_i)$, and $O_{\varpi_3}(N_i)$ being the lower-order remainder terms. Omitting the lower-order remainder terms, consequently, the effect of additive noise to the proposed whitening is approximately characterized by

$$\frac{\Delta\varpi_1}{(\Sigma_{i=1}^4 H_i)^2S^2} \simeq \frac{\left(\frac{N_1}{S}\right)^2+\left(\frac{N_2}{S}\right)^2+\left(\frac{N_3}{S}\right)^2+\left(\frac{N_4}{S}\right)^2}{(\Sigma_{i=1}^4 H_i)^2}, \quad (26)$$

$$\frac{\Delta\varpi_2}{(H_3-H_1)^2S^2} \simeq \frac{\left(\frac{N_1}{S}\right)^2+\left(\frac{N_3}{S}\right)^2}{(H_3-H_1)^2}, \quad (27)$$

$$\frac{\Delta\varpi_2}{(H_4-H_2)^2S^2} \simeq \frac{\left(\frac{N_2}{S}\right)^2+\left(\frac{N_4}{S}\right)^2}{(H_4-H_2)^2}, \quad (28)$$

which implies that the effect of the additive noise on the proposed whitening weighting depends inversely on just the second power of SNR. Therefore, we can expect that the impact of additive noise on the proposed whitening will be less severe under the condition of low SNR, compared to the existing whitening that depends inversely on the fourth power of SNR.

By using Eq. (15), the modified SI features with the whitening weighting can be expressed as

$$I_x^{(W)}(\omega,t) = \frac{1}{W(\omega,t)}Re\{P_0(\omega,t)V_x^\star(\omega,t)\}, \quad (29)$$

$$I_y^{(W)}(\omega,t) = \frac{1}{W(\omega,t)}Re\{P_0(\omega,t)V_y^\star(\omega,t)\}. \quad (30)$$

Another problem with the existing feature extraction using first-order DMAs[25] is that the SI features are further averaged over all the T-F bins. Although the averaging of SI features can significantly reduce the data dimension to facilitate the postprocessing by the support vector machine (SVM), it leads to the complete loss of useful local information on source locations. In this work, because we utilize a CNN for sound source localization, which can handle high-dimensional data and is great for capturing local information, we have thus retained all the T-F bin-wise SI features over the whole T-F domain as the features to the CNN.

2. Enriching SI feature extraction via incorporating subarrays

The above source location feature extraction based on SI estimation utilizes signals received by all the four microphones. Actually, only three microphones are enough to estimate the SI by forming two orthogonal first-order DMAs with a shared microphone. Therefore, we can further decompose the original four-element orthogonal first-order DMAs into four independent three-element subarrays to enrich the SI features via incorporating the redundancies in SI estimation.

Figure 3 shows the decomposed four subarrays, each forming two orthogonal first-order DMAs with shared microphone M_1, M_2, M_3, and M_4, respectively.

For the subarray with the shared microphone M_1 shown in Fig. 3(a), where M_2-M_1 forms the first-order DMA along the direction $3\pi/4$ and M_4-M_1 along the direction $-3\pi/4$, similar to Eqs. (5) and (6), the particle velocity along the direction $-3\pi/4$ and that along the direction $3\pi/4$ at microphone M_1 can be approximately expressed, respectively, as

$$V_{-3\pi/4,M_1}(\omega,t) = \frac{j\sqrt{2}[P_1(\omega,t)-P_4(\omega,t)]}{\omega\rho d}, \quad (31)$$

$$V_{3\pi/4,M_1}(\omega,t) = \frac{j\sqrt{2}[P_1(\omega,t)-P_2(\omega,t)]}{\omega\rho d}. \quad (32)$$

Accordingly, the instantaneous complex SI along directions $-3\pi/4$ and $3\pi/4$ at microphone M_1 is given by

$$I_{-3\pi/4,M_1}(\omega,t) = P_{M_1}(\omega,t)V^\star_{-3\pi/4,M_1}(\omega,t), \quad (33)$$

$$I_{3\pi/4,M_1}(\omega,t) = P_{M_1}(\omega,t)V^\star_{3\pi/4,M_1}(\omega,t), \quad (34)$$

where the sound pressure is estimated by $P_{M_1}(\omega,t) = [P_1(\omega,t)+P_2(\omega,t)+P_4(\omega,t)]/3$.

Similar to the procedures as in Eqs. (29) and (30), the SI features based on the subarray M_2-M_1-M_4 with the proposed whitening weighting given by Eq. (15) can be now formulated as

https://doi.org/10.1121/10.0003445

$$I^{(W)}_{-3\pi/4,M_1}(\omega,t)=\frac{1}{W_{M_1}(\omega,t)}Re\Big\{P_{M_1}(\omega,t)V^{\star}_{-3\pi/4,M_1}(\omega,t)\Big\}, \tag{35}$$

$$I^{(W)}_{3\pi/4,M_1}(\omega,t)=\frac{1}{W_{M_1}(\omega,t)}Re\Big\{P_{M_1}(\omega,t)V^{\star}_{3\pi/4,M_1}(\omega,t)\Big\}, \tag{36}$$

where

$$W_{M_1}(\omega,t)=\Big[|P_{M_1}(\omega,t)|^2+\beta(|V_{-3\pi/4,M_1}(\omega,t)|^2 +|V_{3\pi/4,M_1}(\omega,t)|^2)\Big]^{1/2}. \tag{37}$$

Now we consider the subarray M_3-M_2-M_1 shown in Fig. 3(b), where microphones M_3, M_2, and M_1 form two orthogonal first-order DMAs with the shared microphone M_2, i.e., M_1-M_2 along the direction $-\pi/4$ and M_3-M_2 along the direction $-3\pi/4$. Following similar procedures as above, the whitened SI features along the directions $-3\pi/4$ and $-\pi/4$ can be constructed as

$$I^{(W)}_{-\pi/4,M_2}(\omega,t)=\frac{1}{W_{M_2}(\omega,t)}Re\Big\{P_{M_2}(\omega,t)V^{\star}_{-\pi/4,M_2}(\omega,t)\Big\}, \tag{38}$$

$$I^{(W)}_{-3\pi/4,M_2}(\omega,t)=\frac{1}{W_{M_2}(\omega,t)}Re\Big\{P_{M_2}(\omega,t)V^{\star}_{-3\pi/4,M_2}(\omega,t)\Big\}, \tag{39}$$

where

$$W_{M_2}(\omega,t)=\Big[|P_{M_2}(\omega,t)|^2+\beta(|V_{-3\pi/4,M_2}(\omega,t)|^2 +|V_{-\pi/4,M_2}(\omega,t)|^2)\Big]^{1/2}, \tag{40}$$

with

$$P_{M_2}(\omega,t)=\frac{1}{3}[P_1(\omega,t)+P_2(\omega,t)+P_3(\omega,t)], \tag{41}$$

$$V_{-\pi/4,M_2}(\omega,t)=\frac{j\sqrt{2}[P_2(\omega,t)-P_1(\omega,t)]}{\omega\rho d}, \tag{42}$$

$$V_{-3\pi/4,M_2}(\omega,t)=\frac{j\sqrt{2}[P_2(\omega,t)-P_3(\omega,t)]}{\omega\rho d}. \tag{43}$$

Next, for the subarray M_4-M_3-M_2 shown in Fig. 3(c), where microphones M_4, M_3, and M_2 construct two orthogonal first-order DMAs with the shared microphone M_3, i.e., M_2-M_3 along the direction $\pi/4$ and M_4-M_3 along the direction $-\pi/4$, the proposed whitened SI features along the directions $\pi/4$ and $-\pi/4$ can be expressed as

$$I^{(W)}_{\pi/4,M_3}(\omega,t)=\frac{1}{W_{M_3}(\omega,t)}Re\Big\{P_{M_3}(\omega,t)V^{\star}_{\pi/4,M_3}(\omega,t)\Big\}, \tag{44}$$

$$I^{(W)}_{-\pi/4,M_3}(\omega,t)=\frac{1}{W_{M_3}(\omega,t)}Re\Big\{P_{M_3}(\omega,t)V^{\star}_{-\pi/4,M_3}(\omega,t)\Big\}, \tag{45}$$

where

$$W_{M_3}(\omega,t)=\Big[|P_{M_3}(\omega,t)|^2+\beta(|V_{-\pi/4,M_3}(\omega,t)|^2 +|V_{\pi/4,M_3}(\omega,t)|^2)\Big]^{1/2}, \tag{46}$$

with

$$P_{M_3}(\omega,t)=\frac{1}{3}[P_2(\omega,t)+P_3(\omega,t)+P_4(\omega,t)], \tag{47}$$

$$V_{\pi/4,M_3}(\omega,t)=\frac{j\sqrt{2}[P_3(\omega,t)-P_2(\omega,t)]}{\omega\rho d}, \tag{48}$$

$$V_{-\pi/4,M_3}(\omega,t)=\frac{j\sqrt{2}[P_3(\omega,t)-P_4(\omega,t)]}{\omega\rho d}. \tag{49}$$

Finally, for the subarray M_1-M_4-M_3 shown in Fig. 3(d), where microphones M_1, M_4, and M_3 formulate two orthogonal first-order DMAs with the shared microphone M_4, i.e., M_1-M_4 along the direction $\pi/4$ and M_3-M_4 along the direction $3\pi/4$, the proposed SI features with whitening weighting along the directions $\pi/4$ and $3\pi/4$ are given, respectively, by

$$I^{(W)}_{\pi/4,M_4}(\omega,t)=\frac{1}{W_{M_4}(\omega,t)}Re\Big\{P_{M_4}(\omega,t)V^{\star}_{\pi/4,M_4}(\omega,t)\Big\}, \tag{50}$$

$$I^{(W)}_{3\pi/4,M_4}(\omega,t)=\frac{1}{W_{M_4}(\omega,t)}Re\Big\{P_{M_4}(\omega,t)V^{\star}_{3\pi/4,M_4}(\omega,t)\Big\}, \tag{51}$$

where

$$W_{M_4}(\omega,t)=\Big[|P_{M_4}(\omega,t)|^2+\beta(|V_{\pi/4,M_4}(\omega,t)|^2 +|V_{3\pi/4,M_4}(\omega,t)|^2)\Big]^{1/2}, \tag{52}$$

with

$$P_{M_4}(\omega,t)=\frac{1}{3}[P_1(\omega,t)+P_3(\omega,t)+P_4(\omega,t)], \tag{53}$$

$$V_{\pi/4,M_4}(\omega,t)=\frac{j\sqrt{2}[P_4(\omega,t)-P_1(\omega,t)]}{\rho\omega d}, \tag{54}$$

$$V_{3\pi/4,M_4}(\omega,t)=\frac{j\sqrt{2}[P_4(\omega,t)-P_3(\omega,t)]}{\rho\omega d}. \tag{55}$$

3. Summary of the proposed SI features

Denote the numbers of the frequency bins and frames for the STFT used to estimate the SI features as M and N,

respectively, where M is usually set to be even when using the fast Fourier transform (FFT) to perform the STFT. Then, to summarize, our proposed SI features based on the two orthogonal first-order DMAs and their subarray counterparts can be expressed as the three-dimensional matrix $\mathbf{\Gamma} \in \mathcal{R}^{N\times(M/2-1)\times 10}$ with its entries given by $\mathbf{\Gamma}(:,:,1) = \mathbf{I}_x^{(W)}$, $\mathbf{\Gamma}(:,:,2) = \mathbf{I}_y^{(W)}$, $\mathbf{\Gamma}(:,:,3) = \mathbf{I}_{-3\pi/4,M_1}^{(W)}$, $\mathbf{\Gamma}(:,:,4) = \mathbf{I}_{3\pi/4,M_1}^{(W)}$, $\mathbf{\Gamma}(:,:,5) = \mathbf{I}_{-\pi/4,M_2}^{(W)}$, $\mathbf{\Gamma}(:,:,6) = \mathbf{I}_{-3\pi/4,M_2}^{(W)}$, $\mathbf{\Gamma}(:,:,7) = \mathbf{I}_{\pi/4,M_3}^{(W)}$, $\mathbf{\Gamma}(:,:,8) = \mathbf{I}_{-\pi/4,M_3}^{(W)}$, $\mathbf{\Gamma}(:,:,9) = \mathbf{I}_{\pi/4,M_4}^{(W)}$, and $\mathbf{\Gamma}(:,:,10) = \mathbf{I}_{3\pi/4,M_4}^{(W)}$, where

$$\mathbf{I}_x^{(W)} = \begin{bmatrix} I_x^{(W)}(\omega_1,t_0) & \cdots & I_x^{(W)}(\omega_{M/2-1},t_0) \\ \vdots & \ddots & \vdots \\ I_x^{(W)}(\omega_1,t_{N-1}) & \cdots & I_x^{(W)}(\omega_{M/2-1},t_{N-1}) \end{bmatrix}, \tag{56}$$

with $I_x^{(W)}(\omega_m,t_n)$ given by Eq. (29) and ω_i $(i = 1,\ldots, M/2-1)$ representing the frequency bins of the positive frequencies of the STFT. Note that the SI features at zero frequency are equal to zero and hence contain no source location information. Thus, the SI features at zero frequency are discarded. Moreover, considering the conjugate symmetry property of the Fourier transform, it suffices to utilize only the SI features over the frequency bins of positive frequencies as shown in (56). The remaining matrices $\mathbf{I}_y^{(W)}$, $\mathbf{I}_{-3\pi/4,M_1}^{(W)}$, $\mathbf{I}_{3\pi/4,M_1}^{(W)}$, $\mathbf{I}_{-\pi/4,M_2}^{(W)}$, $\mathbf{I}_{-3\pi/4,M_2}^{(W)}$, $\mathbf{I}_{\pi/4,M_3}^{(W)}$, $\mathbf{I}_{-\pi/4,M_3}^{(W)}$, $\mathbf{I}_{\pi/4,M_4}^{(W)}$, and $\mathbf{I}_{3\pi/4,M_4}^{(W)}$ are similarly defined as Eq. (56), based on Eqs. (30), (35), (36), (38), (39), (44), (45), (50), and (51), respectively.

B. Network architecture

It is known that speech spectrograms have local correlations in the time and frequency domain, and CNNs are able to effectively model those correlations via local connection. Moreover, CNNs can also capture the translational invariance, such as frequency shift due to speaking styles or speaker variations.[32] Therefore, CNNs have been widely used in speech enhancement and speech source localization.[17,20,32] In this work, we also employ CNNs to perform speech source localization.

As shown in Fig. 2(c), the CNN is composed of an input layer, two convolutional layers, two fully connected layers, and an output layer. Each convolutional layer uses 64 convolution kernels with the size of 3×3 to learn local correlations between local T-F regions. Then a batch normalization (BN) layer is used after each convolutional layer to improve the stability of the network and speed up the convergence of the network. The activation function of convolutional layers and fully connected layers is rectified linear units (ReLU).[33] Between the convolutional layer and the fully connected layer and after each fully connected layer, a dropout procedure[34] with rate 0.5 is used to avoid overfitting. The size and number of convolution kernels and the number of nodes in the fully connected layers are shown in Fig. 2(c).

The proposed SI features are first put into convolution layer and the corresponding outputs are determined according to

$$\boldsymbol{X} = f(BN(\mathbf{W}_{c2} * f(BN(\mathbf{W}_{c1} * \mathbf{\Gamma} + \mathbf{b}_{c1})) + \mathbf{b}_{c2})), \tag{57}$$

where $\mathbf{W}_{c1}$ and $\mathbf{W}_{c2}$ refer to the weight of the convolution kernel corresponding to the first and second convolution layer, respectively; $\mathbf{b}_{c1}$ and $\mathbf{b}_{c2}$ stand for an additive bias corresponding to the first and second convolution layer, respectively; and f denotes the ReLU activation function, which is defined as $f(x) = \max(0,x)$.

The full-connected layer combines all the features extracted by the convolution layer to reduce the input two-dimensional feature matrix to a one-dimensional feature vector to facilitate the output layer for classification processing. In the final layer of the network, the softmax activation function[35] is used to perform classification, and the posterior probability for the sound source DOA candidates can be expressed as

$$P_{CNN}(\theta_k|\mathbf{\Gamma}) = \exp(o_k) \Big/ \sum_{k=1}^{K} \exp(o_k), k = 1,2,\ldots,K, \tag{58}$$

where o_k is the output value of the output layer corresponding to the kth class. The final source DOA is estimated by maximizing the posterior probability, i.e.,

$$\hat{\theta} = \arg\max_{\theta_k}(P_{CNN}(\theta_k|\mathbf{\Gamma})), \tag{59}$$

where $\hat{\theta}$ denotes the estimated source DOA.

In the CNN training, the cross-entropy function[36] is used as the loss function, which is given by

$$L = -\sum_{k=1}^{K} z_k \log(P_{CNN}(\theta_k|\mathbf{\Gamma})), \tag{60}$$

where z_k denotes the ground-truth label corresponding to the kth class. We employ the Adam[37] as the optimizer. The initial learning rate is set to be 10^{-3}, and the maximum number of epochs is chosen as 100. Early stopping with a patience of 10 epochs measured on the validation set is also used to prevent overfitting.

IV. SIMULATION EVALUATION

In this section, we compare the proposed sound source localization method with some baseline methods under different acoustic environments, array sizes, and microphone imperfections. The section starts with a description of the baseline methods, evaluation metrics, and simulation setup and then presents the simulation results.

https://doi.org/10.1121/10.0003445

A. Baseline methods

The baseline methods include SI-PHAT-Normal-Redund-LSSVM,[25] which uses the LSSVM as the classifier, and the other three using the CNN as the classifier, i.e., (1) with the SI features following Ref. 25 (SI-PHAT-Normal-Redund-CNN), (2) with the basic SI features based on the full-scale array only (SI-CNN), and (3) with the popularly used features in sound source localization based on the GCC-PHAT (GCC-PHAT-CNN).[16–19] For ease of notation, we briefly summarize these baseline methods as follows:

- *SI-PHAT-Normal-Redund-LSSVM:* This method was proposed in Ref. 25. Here we follow the notation therein. In the method, the SI features are constructed by averaging over all the T-F bins to reduce the data dimension to facilitate the processing by the LSSVM.
- *SI-PHAT-Normal-Redund-CNN:* It is of interest to see the performance of the SI features extracted by directly following Ref. 25 when the CNN is used as the classifier instead of the LSSVM. To make the features adapt to the CNN processing, herein we have dropped the averaging operations during the feature extraction to retain the SI information over the whole T-F domain. The resultant method is referred to as SI-PHAT-Normal-Redund-CNN, following the notation in Ref. 25. Moreover, to achieve better learning effect for the method, the BN layer after each convolution layer in the CNN is removed, and a dropout layer with rate 0.5 is added to the first convolutional layer.
- *SI-CNN:* We also compare our proposed features with the basic SI features, which are extracted based on the full-scale array only without employing whitening weighting, and the resultant method is denoted as SI-CNN.
- *GCC-PHAT-CNN:* Besides the above SI-based features, we will also include the popularly used features in sound source localization based on the GCC-PHAT[16–19] for comparison, and the baseline method using the GCC-PHAT features is termed GCC-PHAT-CNN.

B. Evaluation metrics

To facilitate algorithm evaluation, we use the localization accuracy as performance metrics, which is defined as

$$AR = \frac{N_c}{N_s} \times 100\%, \tag{61}$$

where N_s represents the number of source locations being evaluated, and N_c denotes the number of source locations that are correctly recognized. Herein, a source is considered being correctly localized if the deviation of the estimated DOA from the actual DOA is within $\pm\theta_0$ for a resolution of θ_0.[18]

C. Simulation setup

As shown in Fig. 4, here, we consider a sound source localization task in a rectangular room with the size of

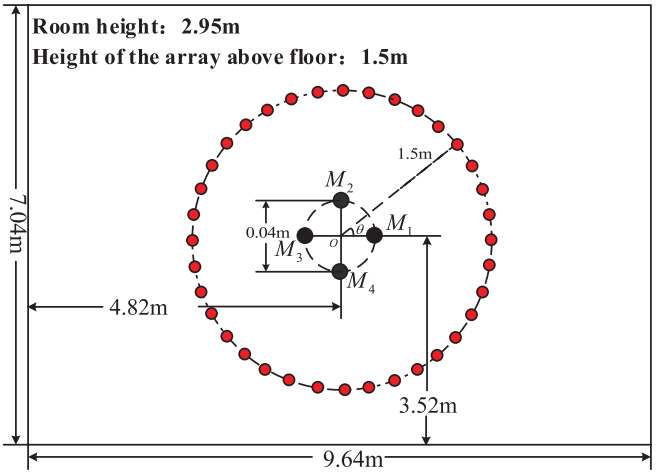

FIG. 4. (Color online) Illustration of the simulation setup. The four black solid dots, located on the inner circle, denote the microphone array. The red solid dots, located uniformly around the outer circle, denote the candidate source locations.

$9.64 \times 7.04 \times 2.95$ m^3. The center of the microphone array, which is constructed by two orthogonal first-order DMAs, is located at $(4.82, 3.52, 1.5)$ m, with the array size $d = 0.04$ m. The distance between the sound source and the center of the microphone array is set to be 1.5 m, and the sound source is at the same height as the array. The RIRs from sound source to microphones are generated using the software RIR GENERATOR,[38] which is based on the well-known image model.[39] The order of sound reflections is set to be the default value defined by the software RIR GENERATOR, which is corresponding to the maximum reflection order automatically calculated by the software given the desired length of RIR,[38] i.e., 3200 in our simulations. The additive noises added on microphones are Gaussian white noise, which is mutually uncorrelated and also is uncorrelated with the sound source signal. The STFT is performed on each frame of 1024 samples with a half-overlapping Hanning window between consecutive frames. For all the SI-based features, the input size of the CNN is 14×511. In addition, the channel C in the CNN is set to be $(2, 6, 10, 10)$, corresponding to (SI-CNN, GCC-PHAT-CNN, SI-PHAT-Normal-Redund-CNN, and the proposed method), respectively. For the LSSVM model, the radial basis function (RBF) is selected as the kernel function, and a tenfold cross-validation is used to tune the regularization parameter and the squared kernel parameter. More details on the training of the LSSVM can be found in Ref. 25.

In the simulations, we have considered two cases with the source location spatial resolutions of 30 and 10°. When the spatial resolution is set to be 30°, the total number of candidate source locations is $K = 12$, which are distributed uniformly from $-170°$ to 160°. The resolution is 10°, and the total number of candidate source locations is $K = 36$, which are distributed uniformly from $-180°$ to 170°. The speech signals are selected from the TIMIT database[40] as the sound source signal, with a duration of 500 ms and sampling frequency of 16 kHz. We randomly select 300

sentences from the training set of the TIMIT database to synthesize the training data, while another 100 sentences are utilized as the validation data. There are in total 6000 and 1000 data points in the training and validation set, respectively. In the testing phase, for each sound source location we randomly select 10 sentences from the test data of the TIMIT database to generate the test set.

D. Simulation results

1. Effect of additive noise

First, we compare the proposed method with the baseline counterparts under different additive noise levels. Here, the SNR varies from 0 to 30 dB with a step size of 5 dB, and the reverberation time is set to be $T_{60} = 300$ ms. In Figs. 5(a) and 5(b), we show the localization accuracy of the various methods as a function of SNR when the spatial resolution is set to be 30° and 10°, respectively.

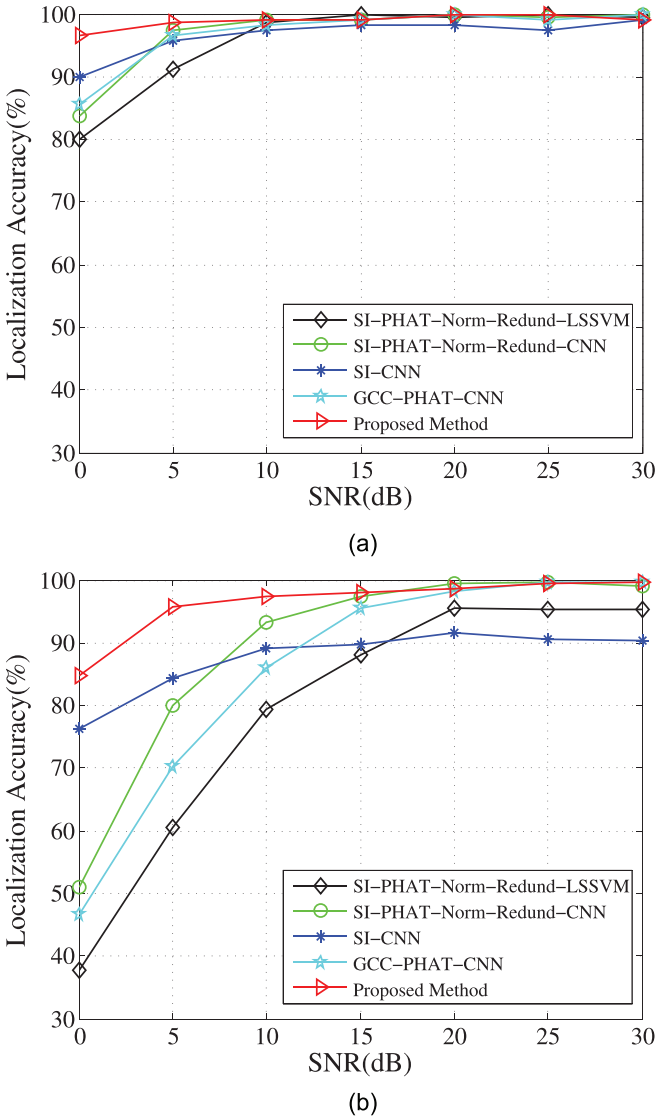

FIG. 5. (Color online) Effect of additive noise on the localization accuracy, with $T_{60} = 300$ ms and spatial resolution of (a) 30° and (b) 10°.

As we can see from the results, when the spatial resolution is 30°, all the CNN- and LSSVM-based methods using SI features have similar recognition performance when the SNR is no less than 10 dB, with the achievable localization accuracy of more than 95%. However, for the more challenging case with the spatial resolution of 10°, as shown in Fig. 5(b), the performance of SI-PHAT-Normal-Redund-LSSVM degrades significantly with the decreasing of SNR. This indicates that SI-PHAT-Normal-Redund-LSSVM fails to work at low SNR conditions with a higher spatial resolution requirement. In contrast, SI-PHAT-Norm-Redund-CNN shows better robustness than SI-PHAT-Normal-Redund-LSSVM, because the SI features across all T-F bins are used and hence more detailed information on source location has been captured, and also the CNN has a more powerful learning ability than the LSSVM with the RBF kernel.[25] Regarding the SVM method, we would like to point out that other kernels may be used instead of the RBF kernel to improve the sound localization performance. However, this is beyond the scope of this work. Nevertheless, because the PHAT weighting suffers in the presence of strong additive noise, as discussed in Sec. III A 1, SI-PHAT-Norm-Redund-CNN is much poorer than the proposed method in terms of the localization accuracy at low SNRs. When SNR = 0 dB with the spatial resolution of 10°, the localization accuracy of the proposed method is 84.72%, while that of SI-PHAT-Norm-Redund-CNN is just 51.11%. This demonstrates that the proposed whitening weighting by decoupling the correlation between the sound pressure and particle velocity components indeed performs better than the existing PHAT-based counterpart under low SNR environments.

For GCC-PHAT-CNN that uses the popular GCC-PHAT features, we can see from the simulation results that it is also more sensitive to additive noise compared with the proposed method and performs rather poorly especially for the case with the higher spatial resolution setting. For SNR = 0 dB with the spatial resolution of 10°, the localization accuracy of GCC-PHAT-CNN is as low as 46.67%. Besides the reason that the PHAT weighting becomes highly sensitive to the additive noise when the SNR is low, another reason is that the microphone array used is small-sized, which has limited the performance of the GCC-PHAT features-based method, as will be shown clearly below in the numerical analysis regarding the effect of the array size.

Recall that, in the proposed SI feature extraction, we propose to use a modified whitening weighting by decoupling the correlation between the sound pressure and particle velocity components to overcome the problem with the existing PHAT-based counterpart. In the modified whitening weighting, we have introduced a varying trade-off parameter β to compensate for the large difference between the power spectra of sound pressure and particle velocity components. Here, we present a numerical example to show the effect of β. In Fig. 6, the localization accuracy of the proposed method under different values of β is presented as a function of SNR, where the spatial resolution of source locations is set to be 10°.

https://doi.org/10.1121/10.0003445

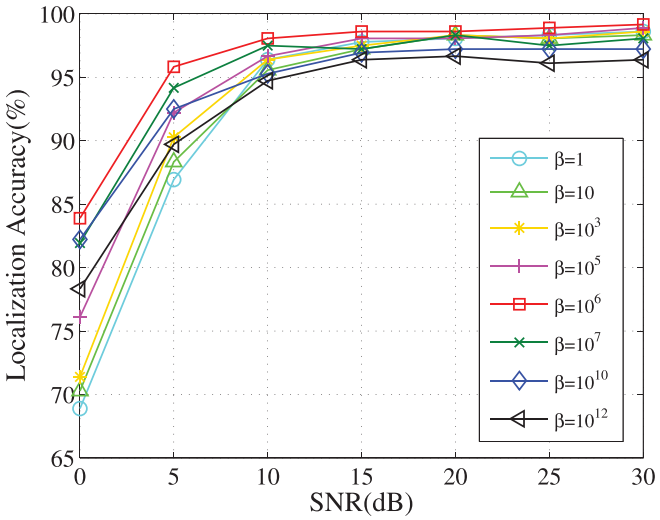

FIG. 6. (Color online) Effect of the trade-off parameter β on the localization accuracy of the proposed method, where $T_{60} = 300$ ms and the spatial resolution is 10°.

It can be seen from Fig. 6 that the best localization performance is achieved around $\beta = 10^6$. Deviation of β from 10^6 will lead to increased sensitivity to the additive noise. This is because the power spectrum of the particle velocity components estimated using the first-order DMAs is far less than that of the estimated sound pressure, and their difference in this simulation example is up to 6 orders of magnitude. Thus, if the β chosen is too small, the contribution of the orthogonal particle velocity components will be ignored in the whitening weighting. On the contrary, β should not be set to be much larger than $\beta = 10^6$; otherwise, the contribution of the sound pressure will also be ignored in the whitening weighting.

2. Effect of room reverberation

In the second example, we compare the effect of room reverberation on the various source localization methods. Figures 7(a) and 7(b) show the localization accuracy of the various methods with the spatial resolution of 30° and 10°, respectively. Here, the reverberation time T_{60} varies from 200 to 800 ms with a step size of 50 ms, and the SNR is set to be 10 dB.

From the simulation results, we can see that although some baseline methods, i.e., SI-PHAT-Norm-Redund-LSSVM and GCC-PHAT-CNN, may produce comparable performance to the proposed method when the spatial resolution is 30°, when the spatial resolution is improved to 10°, all the baseline methods degrade to be obviously worse than the proposed method under various reverberant environments. Generally speaking, the LSSVM-based method, SI-PHAT-Norm-Redund-LSSVM, performs poorer than the CNN-based counterparts, particularly under the condition of high spatial resolution. This further implies that the local information on source locations in SI feature extraction, which is lost in SI-PHAT-Norm-Redund-LSSVM, is helpful for improving the spatial resolution. Note that SI-CNN does

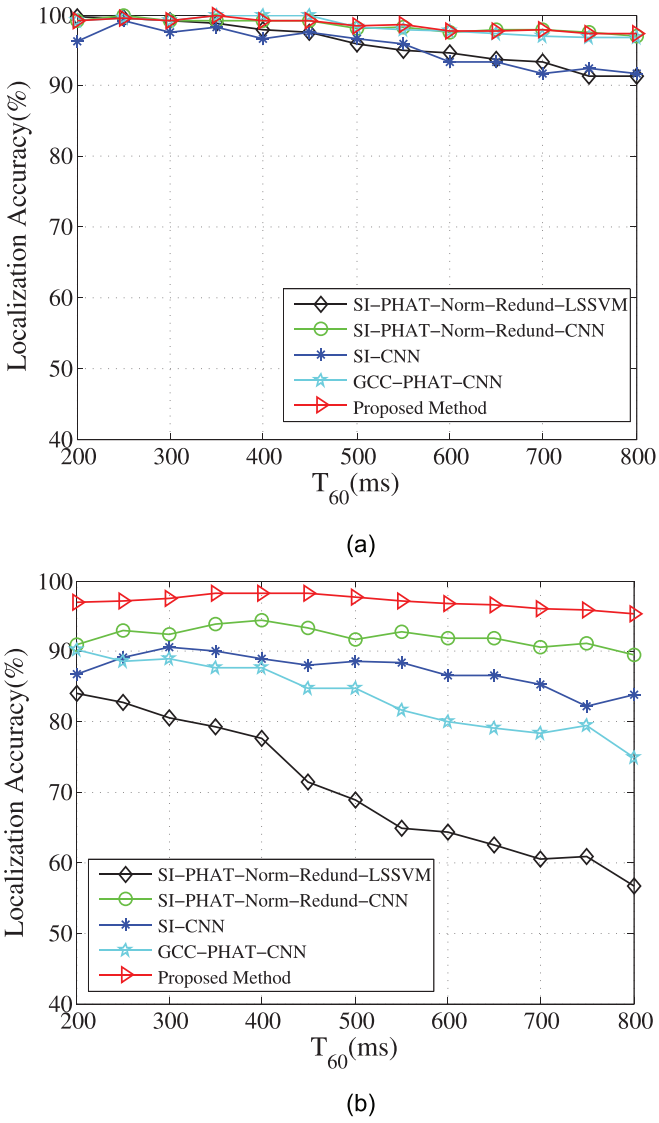

FIG. 7. (Color online) Effect of room reverberation on the localization accuracy, with SNR = 10 dB and the spatial resolution of (a) 30° and (b) 10°.

not conduct the whitening weighting to combat room reverberation and also has not incorporated the redundancies in SI estimation to enrich the feature extraction, when compared to its CNN-based counterparts using the SI features. Thus, SI-CNN shows the worst performance in the reverberant environments among the CNN-based methods using the SI features. Regarding GCC-PHAT-CNN, which uses the conventional GCC-PHAT features, we can see from Fig. 7(a) that when the spatial resolution is low, it can achieve satisfactory performance comparable to the proposed method. When the spatial resolution is increased, however, it degrades significantly with the increasing of reverberation time and becomes much worse than its counterparts using the SI features, as shown in Fig. 7(b).

3. Effect of array size

Now we discuss the effect of array size on the performance of the different methods. Figure 8 shows the

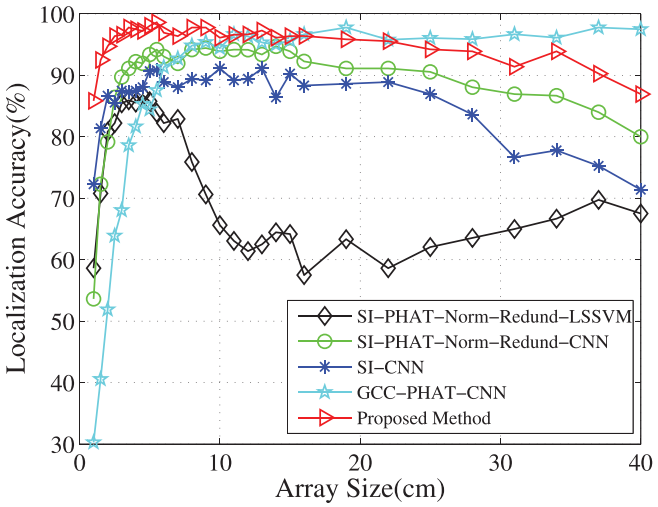

FIG. 8. (Color online) Effect of array size on the localization accuracy, where $T_{60} = 300$ ms, SNR = 10 dB, and the spatial resolution is set to be 10°.

localization accuracy of the various methods as a function of the size of the microphone array d, where the spatial resolution of source locations is chosen to be 10°. Here, the array size d varies from 1 to 40 cm, with $T_{60} = 300$ ms and SNR = 10 dB.

From the simulation results in Fig. 8, we can see that when the SI information is used as features for source localization, no matter which kind of classifiers are used, the array size shows a similar effect on all SI-based methods. That is, the localization performance deteriorates when the array size is either too small or too large. One reason is that the SI features are based on the finite difference of the acoustic pressure signals for the particle velocity approximation using the first-order DMAs, i.e., according to Eq. (4). As a result, when the array size is increased, the approximation error will be accordingly increasing, thus leading to poor SI estimation. On the other hand, if the array size is set too small, the DMAs will suffer from poor noise sensitivity, especially at low frequencies.[41] According to extensive simulation results, the optimal array size range is 2–5.5 cm for the proposed method. For GCC-PHAT-CNN, we can see from the results that it is just suitable for the case with a relatively large array size. For a small array size, as used in our simulations with $d = 4$ cm, its localization accuracy may be much worse than the SI-based counterparts.

4. Effect of microphone mismatches

It is known that small-sized microphone arrays are usually sensitive to microphone mismatches, such as microphone gain and phase errors. In this example, we compare the performance of the various methods in the presence of microphone gain and phase errors.

Suppose that the microphone gain and phase errors are unknown and bounded, respectively, by

$$\varepsilon \in \lambda[-0.1, 0.1], \quad \psi \in \lambda[-5^\circ, 5^\circ], \tag{62}$$

where ε denotes the microphone gain error, ψ refers to the microphone phase error, and λ represents the scale parameter used to control the error ranges. In the simulations, the microphone gain and phase errors of each microphone are randomly selected within the error range given by Eq. (62).

Figure 9 show the localization accuracy of the various methods with λ varying from 0 to 1 with a step size of 0.2, i.e., with the microphone mismatches increasing, where the reverberation time $T_{60} = 300$ ms, SNR = 10 dB, and the spatial resolution of source locations is set to be 10°. From Fig. 9, we can see that, compared with the CNN-based methods, the LSSVM-based method, SI-PHAT-Norm-Redund-LSSVM, degrades significantly with the increasing of microphone mismatches. This may be because the LSSVM essentially belongs to a shallow neural network, which contains just one layer of nonlinear feature transformation, so its generalization ability is limited and cannot effectively learn the information about the sound source locations from the features corrupted by mismatch errors, leading to a poor ability in error tolerance. In contrast, the CNN is a deep architecture with more hidden layers, which can automatically extract the most effective features during the training process to improve the error tolerance ability of the neural network. Therefore, the localization performance of the four CNN-based methods is basically insensitive to microphone imperfections. Moreover, due to the effectiveness of our feature extraction, the localization accuracy of the proposed method always maintains above 95% and shows superior performance among all the CNN-based methods in the presence of microphone mismatches.

V. REAL-WORLD EXPERIMENTAL RESULTS

In this section, some real-world experiments were conducted in a room environment to evaluate the performance of the studied methods. The setup of the experiments is shown in Fig. 10(a). The size of the room is 9.64 × 7.04 ×

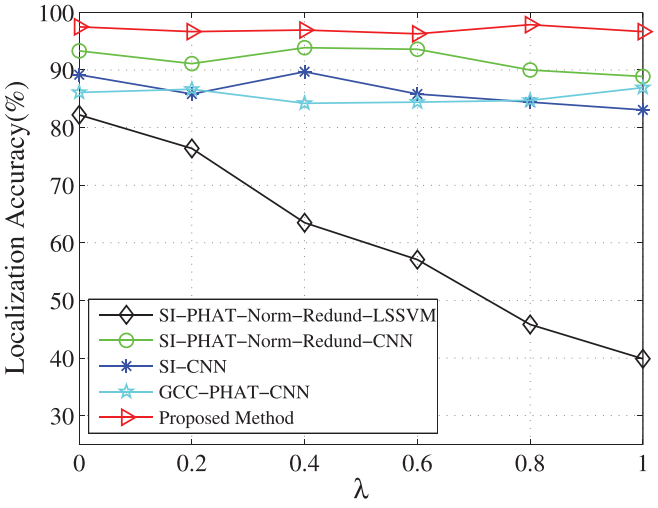

FIG. 9. (Color online) Effect of microphone mismatches on the localization accuracy, where λ is the scale parameter used to control the range of microphone mismatches as given in Eq. (62), $T_{60} = 300$ ms, SNR = 10 dB, and the spatial resolution is set to be 10°.

https://doi.org/10.1121/10.0003445

(a) (b)

FIG. 10. (Color online) (a) Photograph of the setup of the real-world experiments. (b) Photograph of the microphone array used in the experiments.

2.95 m^3, and the reverberation time of the room was measured around 300 ms. We used a microphone array similar to that in the above simulations with a size of 4 cm, which consists of four $\frac{1}{2}$-inch microphones forming two orthogonal first-order DMAs, as shown in Fig. 10(b). The microphone array was placed around the center of the room, with a height of 1.5 m above the floor. We used a loudspeaker as the sound source, with a distance of 1.5 m from the center of the microphone array. The NI's USB-4432 data acquisition device with a 24-bit resolution was employed with a sampling frequency of 16 kHz. In the experiments, we have considered two cases with different spatial resolution, i.e., 30° and 20°.

Unlike Ref. 25, which employed the real-world experimental data for the training of a neural network, here we instead used the same simulated data as in the above simulation examples for the training of the CNN and LSSVM (note that in the simulations, we have used a room with the same size as that in the real-world experiments). By doing so, we can examine the capability of the various methods in dealing with the mismatch between the training and test conditions, and also the repetition of laborious experiments will not be required to collect a large amount of data for neural network training. The parameter settings of the CNN and LSSVM for the real-world experiments are the same as those in the above simulations. In the source localization phase, 20 different speech signals were randomly selected at each location, which results in a total of 240 and 360 test data points corresponding to the spatial resolution of 30° and 20°, respectively.

Tables I and II show the localization accuracy of the various methods when the spatial resolution is 30° and 20°, respectively. As can be seen from the results, generally speaking, the real-world experimental results are well consistent with the simulation results presented in Sec. IV. Although some baseline methods can achieve comparable performance to the proposed method for the case with low spatial resolution, all of them deteriorate dramatically when the spatial resolution becomes higher. In contrast, the proposed method is less sensitive to the spatial resolution alteration and shows the best performance among all the studied methods in the practical reverberant environment.

TABLE I. Localization accuracy of the various methods in the real experiments, where the spatial resolution is 30°. The results are shaded, with white being the lowest performance and black being the highest.

DOA (deg)	SI-PHAT-Norm-Redund-LSSVM (%)	SI-PHAT-Norm-Redund-CNN (%)	SI-CNN (%)	GCC-PHAT-CNN (%)	Proposed method (%)
−170	90	85	85	90	95
−140	85	90	80	95	100
−110	100	100	90	90	95
−80	70	100	80	100	100
−50	100	90	100	90	95
−20	90	85	80	95	95
−10	95	100	75	90	95
40	100	100	85	90	100
70	100	100	80	100	95
100	100	100	85	95	100
130	50	80	45	95	90
160	100	100	80	95	100
Average	90	94.16	80.41	93.75	96.67

TABLE II. Localization accuracy of the various methods in the real experiments, where the spatial resolution is 20°. The results are shaded, with white being the lowest performance and black being the highest.

DOA (deg)	SI-PHAT-Norm-Redund-LSSVM (%)	SI-PHAT-Norm-Redund-CNN (%)	SI-CNN (%)	GCC-PHAT-CNN (%)	Proposed method (%)
−170	70	85	70	60	90
−150	35	85	55	65	95
−130	90	90	25	90	90
−110	90	75	70	80	85
−90	50	90	35	95	100
−70	50	70	80	75	95
−50	55	85	70	55	90
−30	55	75	60	60	95
−10	100	100	50	70	100
10	30	85	75	80	95
30	70	80	60	85	100
50	45	90	60	80	90
70	95	95	75	90	95
90	100	90	40	80	100
110	60	75	35	85	90
130	35	80	80	80	90
150	95	90	90	75	100
170	15	75	45	40	85
Average	63.33	84.16	59.72	74.72	93.61

For SI-PHAT-Norm-Redund-LSSVM, the performance is inferior to that of its counterpart, SI-PHAT-Norm-Redund-CNN, especially for the case with a high spatial resolution of 20°. Similar to the findings in the simulation results, this demonstrates that the local information on source locations in SI feature extraction as employed in SI-PHAT-Norm-Redund-CNN is helpful to improve the localization performance in practical environment. Among the CNN-based methods using SI features, SI-CNN performs the poorest, which further confirms that the whitening weighting and use of redundancies in SI estimation are indispensable for constructing robust source location features. Moreover, the proposed method has shown better performance than SI-PHAT-Norm-Redund-CNN, with the average localization accuracy increased by nearly 10% for the more challenging case with spatial resolution of 20°. As also revealed in simulations, this again implies that the proposed whitening weighting scheme is more robust than its existing PHAT weighting counterpart. Regarding the CNN-based approach using the popular GCC-PHAT features, we can also see that the GCC-PHAT feature extraction is not applicable to the small-sized array because it may lead to rather poor performance under the condition of a high spatial resolution. For the real-world experiments with the spatial resolution of 20°, the average localization accuracy of GCC-PHAT-CNN is almost 20% below that of the proposed method.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an indoor sound source localization algorithm assisted by CNN for a small-sized microphone array consisting of two orthogonal first-order DMAs. Particularly, an improved sound location feature extraction scheme is proposed based on our previously proposed SI features, which is shown to be robust against additive noise, room reverberation, and microphone mismatches. Unlike the existing SI features, which suffer from the complete loss of useful local information on source location over the T-F domain, the proposed SI features have retained all the T-F bin-wise SI features. Moreover, the sensitivity of the existing SI features to additive noise is also analyzed. It is revealed that the existing SI features are sensitive to additive noise due to the correlation between the sound pressure and particle velocity components in the whitening weighting. And an improved SI feature extraction scheme is then proposed by decoupling this correlation relationship in the whitening weighting construction.

Regarding the GCC-PHAT features, which are widely adopted in the literature for sound localization, our analysis on the effect of array size shows that the GCC-PHAT features are not suitable for small-sized microphone arrays, such as the array used in the paper with just a 4-cm diameter. In contrast, the proposed SI features perform much better for the small-sized array, especially under the condition of high spatial resolution of sound locations. The effect of microphone mismatches, which is a main concern for the design of small-sized arrays, is also analyzed. It is shown that the proposed method is interestingly less sensitive to varying sensor imperfections and hence is promising in practical applications. Simulation and real-world experimental results have consistently demonstrated the superior performance of the proposed sound localization method compared to its existing counterparts.

Finally, we would like to point out some future work. In the present study, we have studied the performance of the proposed localization model when the training conditions are similar to its targeted applied scenario. However, we may want the localization model to be able to work in a

room with a different size from those in the training process. To this end, it is of interest to study the generalizability of the proposed localization model by including more training data with various room sizes. In addition, the present work deals with the localization problem for a single speech source. Other future work will be to extend the proposed method to address localization of multiple speech sources. This is possible by using the so-called W-disjoint orthogonality property of speech signals,[42] i.e., at most only one speech source will be active at each T-F bin, which has been widely employed in speech source separation and multiple speech source localization.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their constructive comments to improve the quality and presentation of the paper. This research 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.

[1]H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, Germany (April 21–24, 1997), pp. 187–190.
[2]L. Calmes, G. Lakemeyer, and H. Wagner, "Azimuthal sound localization using coincidence of timing across frequency on a robotic platform," J. Acoust. Soc. Am. **121**, 2034–2048 (2007).
[3]A. Archer-Boyd, W. Whitmer, W. Brimijoin, and J. Soraghan, "Biomimetic direction of arrival estimation for resolving front-back confusions in hearing aids," J. Acoust. Soc. Am. **137**(5), EL360–EL366 (2015).
[4]C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," IEEE Trans. Multimedia **10**(3), 538–548 (2008).
[5]N. Roman, D. Wang, and G. J. Brow, "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**(4), 2236–2252 (2003).
[6]C. Busso, S. Hernanz, C. W. Chu, S. Kwon, S. Lee, P. G. Georgiou, I. Cohen, and S. Narayanan, "Smart room: Participant and speaker localization and identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA (March 23, 2005), pp. 1117–1120.
[7]C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (1976).
[8]Y. T. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," IEEE Trans. Speech Audio Process. **9**(8), 943–956 (2001).
[9]J. Dmochowski, J. Benesty, and S. Affes, "Fast steered response power source localization using inverse mapping of relative delays," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV (March 31–April 4, 2008). pp. 289-292.
[10]A. Martia, M. Cobos, J. J. Lopez, and J. Escolano, "A steered response power iterative method for high-accuracy acoustic source localization," J. Acoust. Soc. Am. **134**, 2627–2630 (2013).
[11]S. Argentieri and P. Danes, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Diego, CA (October 29–November 2, 2007). pp. 2009–2014.
[12]R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," IEEE Trans. Acoust. Speech Signal Process. **37**(7), 984–995 (1989).
[13]J. Benesty, J. Chen, and and Y. Huang, *Microphone Array Signal Processing* (Springer, New York, 2008).
[14]M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, C. A. Deledalle, and W. Li, "Machine learning in acoustics: Theory and applications," J. Acoust. Soc. Am. **146**(5), 3590–3628 (2019).
[15]H. Kayser and J. Anemuller, "A discriminative learning approach to probabilistic acoustic source localization," in *Proceedings of the 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France (September 8–11, 2014), pp. 99–103.
[16]X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, South Brisbane, Australia (April 19–24, 2015), pp. 2814–2818.
[17]X. Yue, G. Qu, B. Liu, and A. Liu, "Detection sound source direction in 3D space using convolutional neural networks," in *Proceedings of the First International Conference on Artificial Intelligence for Industries (AI4I)*, Laguna Hills, CA (September 26–28, 2018), pp. 81–84.
[18]Q. Li, X. Zhang, and H. Li, "Online direction of arrival estimation based on deep learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada (April 15–20, 2018), pp. 2616–2620.
[19]Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," IEEE Trans. Ind. Electron. **65**(8), 6403–6413 (2018).
[20]P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (*ICASSP), New Orleans, LA (March 5–9, 2017), pp. 6125–6129.
[21]Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Hyderabad, India (September 2–6, 2018). pp. 322–326.
[22]Z. Q. Wang, X. Zhang, and D. L. Wang, "Robust speaker localization guided by deep learning based time-frequency masking," IEEE/ACM Trans. Audio Speech Lang. Process. **27**(1), 178–188 (2019).
[23]P. Pertilä and M. Parviainen, "Time difference of arrival estimation of speech signals using deep neural networks with integrated time-frequency masking," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK (May 12–17, 2019), pp. 436–440.
[24]E. Ozanich, P. Gerstoft, and H. Niu, "A feedforward neural network for direction-of-arrival estimation," J. Acoust. Soc. Am. **147**(3), 2035–2048 (2020).
[25]Y. Li and H. Chen, "Reverberation robust feature extraction for sound source localization using a small-sized microphone array," IEEE Sens. J. **17**(19), 6331–6339 (2017).
[26]J. Escolano, J. M. Perez-Lorenzo, N. Xiang, M. Cobos, and J. J. López, "A Bayesian inference model for speech localization," J. Acoust. Soc. Am. **132**(3), 1257–1260 (2012).
[27]J. Escolano, N. Xiang, J. M. Perez-Lorenzo, M. Cobos, and J. J. Lopez, "A Bayesian direction-of-arrival model for an undetermined number of sources using a two-microphone array," J. Acoust. Soc. Am. **135**(2), 742–753 (2014).
[28]I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).
[29]F. Jacobsen, "Sound intensity," in *Springer Handbook of Acoustics*, edited by T. D. Rossing (Springer, New York, 2014), pp. 1093–1114.
[30]F. J. Fahy, *Sound Intensity*, 2nd ed. (Spon Press, London, 1995).
[31]H. Hacihabiboğlu, "Theoretical analysis of open spherical microphone arrays for acoustic intensity measurements," IEEE/ACM Trans. Audio Speech Lang. Process. **22**(2), 465–476 (2014).
[32]Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," IEEE/ACM Trans. Audio Speech Lang. Process. **24**(12), 2263–2276 (2016).
[33]V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International*

Conference on Machine Learning (ICML), Haifa, Israel (June 21–24, 2010), pp. 807–814.
[34]N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res. **15**(1), 1929–1958 (2014).
[35]R. A. Dunne and N. A. Campbell, "On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function," in *Proceedings of the 8th Australian Conference on Neural Networks (ACNN)* (1997), pp. 181-185.
[36]J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," IEEE Trans. Inform. Theory **26**(1), 26–37 (1980).
[37]D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980 (2014).
[38]E. A. P. Habets, "Room Impulse Response (RIR) Generator" (2016), https://github.com/ehabets/RIR-Generator Last viewed January 13, 2021.
[39]J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**(4), 943–950 (1979).
[40]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-ROMs," National Institute of Standards and Technology, Gaithersburg, MD, NIST Interagency/Internal, Report No. 4930 (1993).
[41]J. Benesty and J. Chen, *Study and Design of Differential Microphone Arrays* (Springer, Berlin, 2013).
[42]O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004).