Acoustic Source Localization in the Circular Harmonic Domain Using Deep Learning Architecture

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

***Abstract*—The problem of direction of arrival (DOA) estimation with a circular microphone array has been addressed with classical source localization methods, such as the model-based methods and the parametric methods. These methods have an advantage in estimating the DOAs in a blind manner, i.e. with no (or limited) prior knowledge about the sound sources. However, their performance tends to degrade rapidly in noisy and reverberant environments or in the presence of sensor array limitations, such as sensor gain and phase errors. In this paper, we present a new approach by leveraging the strength of a convolutional neural network (CNN)-based deep learning approach. In particular, we design new circular harmonic features that are frequency-invariant as inputs to the CNN architecture, so as to offer improvements in DOA estimation in unseen adverse environments and obtain good adaptation to array imperfections. To our knowledge, such a deep learning approach has not been used in the circular harmonic domain. Experiments performed on both simulated and real-data show that our method gives significantly better performance, than the recent baseline methods, in a variety of noise and reverberation levels, in terms of the accuracy of the DOA estimation.**

***Index Terms*—microphone array signal processing, circular harmonic, acoustic source localization, deep learning architecture, convolutional neural network (CNN).**

I. Introduction

ACOUSTIC source localization using sensor arrays is an active area of research in speech signal processing [1], with various practical applications. For example, in home assisted living, the ability to localize a speaker in daily environments is important for a voice-based interface such as Amazon Echo and Google Home [2]; in autonomous driving, sound source localization is useful for autonomous vehicles in perceiving their surrounding environment [3]; in audio surveillance systems, the spatial information derived from the measurements of microphones can be used as a fundamental monitoring unit of drones [4] and robots [5]; in human-computer interaction, the directions of sound sources are essential for enhancing the sources of interest or for reproducing the acoustic scene with the presence of such sources [6], among many others [7]–[9].

Direction of arrival (DOA) estimation methods in the literature can be broadly classified into four categories. The first one is based on the time difference of arrival (TDOA) of sound sources, e.g. by exploiting the Generalized Cross Correlation PHAse Transform (GCC-PHAT), which was originally designed to localize a single sound source with the highest signal-to-noise ratio (SNR) in an environment [10], and then extended to localize multiple sources with various SNRs [11]. While in environments with strong reverberation and directional or diffuse noise, the summation of the GCC-PHAT coefficients would exhibit high peaks from interference sources [12]. The second category is represented by the subspace-based approaches, including the popular methods e.g. multiple signal classification (MUSIC) [13] and estimation of signal parameters via rotational invariance techniques (ESPRIT) [14]. Subspace methods can be applied with different array types and produce high resolution DOA estimates for multiple narrow-band sources. They are generally robust to diffuse noise, whereas they tend to be sensitive to directional noise sources and room reverberations [15]. The reason is that the noise subspace constructed from the eigenvectors corresponding to the smallest eigenvalues of noisy speech covariance matrices may not be the true noise subspace [12]. The third are beamforming based approaches such as steered response power with phase transform (SRP-PHAT) [16] and minimum variance distortionless response (MVDR) [17] methods. Owing to their limitations on spatial resolution, these methods may fail to localize the speakers that are close to each other [18]. The fourth methods are based on sound intensity vectors, which determine the magnitude and direction of the transport of acoustic energy, related to the DOA of a sound wave [19]. Unfortunately, it is difficult to measure particle velocity, although attempts have been made to use the finite difference method with two microphone arrays [20].

In the past several years, modal signal processing using sensor array has received increasing attention [21], since it can provide a frequency-invariant eigenbeam that is useful for localizing wideband source without a narrowband assumption underlying the traditional signal model [22]. The authors of [23] developed the time-frequency circular harmonic beamforming (TF-CHB), which was reported to achieve better DOA resolution than the eigenbeam (EB)-ESPRIT under a high level of reverberation and noise. In [19], a method using the pseudointensity vector (PIV) is designed for the localization of a single source, which uses sound field information with low order spatial information. In [24], MUSIC with direct-path dominance (DPD) test is used to improve source localization in highly reverberant environments by exploiting the sparsity of speech in the TF domain. Furthermore, in our earlier work [25], [26], model and parametric methods for circular harmonic DOA estimation have been proposed to enhance the

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61971219, and in part by the Key Laboratory of System Control and Information Processing, Ministry of Education under Grant Scip201802. (*Corresponding author: H. Chen*)

K. SongGong and H. Chen are with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China and also with the Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai 200240, China.(email: sgkk@nuaa.edu.cn, hwchen@nuaa.edu.cn).

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

localization robustness in adverse environments.

Although the aforementioned circular harmonic DOA estimation approaches have achieved certain degree of success in a variety of acoustic conditions, they are still limited in one or more of the following aspects. First, their performance degrades in adverse environments with high-level background noise e.g. SNR below 0 dB, or high-level room reverberation with reverberation time in RT60 greater than 700 ms. Specifically, since mode strengths are sensitive to background noise and reverberation, the performance of circular harmonic based methods degrades in such environments. Second, the characteristics of the array, particularly the small-sized array, may affect the localization performance or limit its practical applications, for instance, the number of transducers, the radius, and whether the sensors are mounted on a scatter such as a rigid cylinder or a sphere. In the TF-CHB method, the localization performance can be improved by increasing the number of sensors and the radius of the array simultaneously as this increases the maximum order that can be used. Third, low accuracy in mismatched conditions. Owing to the lack of adaptation to various array imperfections, uncertainties would be introduced to the DOA estimation, which may lead to negative impacts on the performance. Fourth, unreliable performance in real acoustic environments. The reason is that accurate localization would require tuning of model parameters which may not be easy to achieve in a complex acoustic environment. Here we aim to develop methods that could potentially mitigate these issues.

Recently, deep learning based methods have been proposed for DOA estimation of acoustic sources. In general, these methods involve a feature extraction step from the acquired speech signals, followed by a deep neural network (DNN) that is trained to map the features from the microphone signals to the DOAs of the sources. In Xiao *et al*. [27], the authors proposed a robust DOA estimation method based on multilayer perceptron neural network, with GCC-PHAT from pairs of microphone signals as the input feature, and the results demonstrated its effectiveness against low level noise and strong reverberations. In Takeda *et al*. [28], similar to the computations involved in the MUSIC method for localization, the eigenvalue decomposition of the spatial correlation matrix was performed to get the eigenvectors corresponding to the noise subspace, and these vectors were provided as input to a DNN. However, the experimental results showed that this method is sensitive to reverberations. In [29], [30], convolutional neural network (CNN) based method was proposed, in which phase component of short time Fourier transform (STFT) was used as feature of CNN. The CNN-based method showed robustness to noise and small perturbations in sensor positions. In Adavanne *et al*. [31], [32], the first-order ambisonic (FOA) signals were applied as inputs to a stacked convolutional and recurrent neural network (CRNN) for localization, which achieved promising results. In [12], [33], the bidirectional long short-term memory (BLSTM) neural network is used for estimating the TF masks (i.e., ideal ratio mask (IRM) and phase sensitive mask (PSM)) at each microphone channel, and only using TF units for multichannel localization. This approach uses deep learning for TF unit level classification or regression for robust localization. Generally speaking, when compared with the conventional DOA estimation counterparts mentioned above, the learning-based methods are data-driven and offer several advantages, for example, (i) they can adapt to diverse unseen acoustic scenarios, (ii) they tend to be more robust against different noise and reverberation levels, and (iii) they do not rely on prior assumptions about array geometries, (e.g., including the condition of array imperfections [34]).

In a recent work [25], we proposed a robust localization method that incorporated circular harmonic pseudointensity vector through joining the least-squares decomposition and the spatial processing. In [26], we presented a multi-speaker localization method in the circular harmonic domain based on the acoustic holography beamforming technique and the Bayesian nonparametrics method. These methods are all based on model and parametric methods, where several model parameters need to be tuned accordingly when they are applied to the practical applications. This may result in unstable DOA estimation in diverse environments and limit their flexibility in these applications. In this paper, we study the use of deep learning based method for acoustic source localization in the circular harmonic domain, which, to our knowledge, has not been done in the literature.

Our novel contributions are on the design of new features for the learning of deep models based on circular harmonic analysis in far-field wave propagation. The first feature is obtained by constructing the matrix of the equalized circular harmonics via using the data model of circular harmonics representation and the related equalization coefficient vector, and using its real and imaginary components. The second feature is obtained by simultaneously utilizing the magnitude and phase of circular harmonic modes as features to improve the localization accuracy, as the phase information is a key factor in DOA estimation. The third feature is built on the second by making use of the energy of zero mode strength to form the circular harmonic enhanced function and employing this function to produce the circular harmonic enhanced modes magnitude and phase as features, which can reduce the impact of noise and room reverberation and further improve the DOA performance.

Based on these features, we present a CNN architecture for robust indoor DOA estimation, by treating the DOA estimation as a classification problem, as in a previous work [29], where each discretized DOA corresponding to a class is recognised with the CNN model. However, DOA estimation can also be treated as a regression problem as in [35]. In our work, the focus is on analyzing the impact of the features based on circular harmonics, therefore, we choose an existing CNN architecture [29] to build the model for classifying the DOA of the sources. Furthermore, our proposed deep learning method is suitable for most circular arrays. Simulations and real-data experiments show the superior performance of our proposed method in noisy and reverberant environments when compared with the existing deep learning methods, and also exhibit good adaptation and robustness to the variations in array and unseen scenarios.

The remainder of this paper is structured as follows. Section II provides a problem formulation. Section III reviews the basic CHB approach in detail. Section IV presents three novel circular harmonic features. Section V shows the practical systems for DOA estimation, which contain two aspects, i,e., CNN network architecture and block diagram of the proposed algorithm. The simulation and real-data experimental results are discussed in Section VI. Finally, the conclusions are drawn in Section VII.

II. Problem Formulation

Consider a sound source with azimuth angle ϕ_s and elevation angle η_s, impinging on a uniform circular sensor array consisting of M omnidirectional sensors, as shown in Fig. 1. The geometric center of the uniform circular array is chosen as the origin of the coordinate system, the radius of the array is r (the diameter $d = 2r$), and the azimuth angle of each sensor is θ_m, namely

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

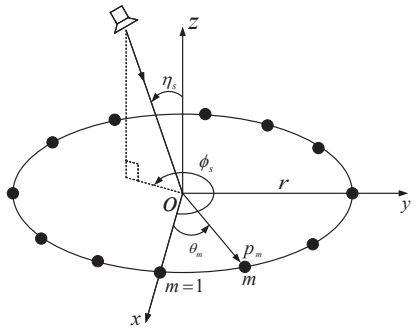


Fig. 1. Configuration of the uniform circular sensor array.

The signal received at the mth sensor can be modeled as

$$p_m(\tilde{t}) \quad = \quad h_m(\tilde{t}) * s(\tilde{t}) + v_m(\tilde{t}), \tag{2}$$

where $s(\tilde{t})$ is the sound source signal, $h_m(\tilde{t})$ is the room impulse responses (RIRs) from the source to the mth sensor, $\otimes$ is the convolution operator, $v_m(\tilde{t})$ is the background noise, and $\tilde{t}$ is a discrete time index.

In the STFT domain, (2) can be transformed to

$$P_m(k,t) \quad = \quad H_m(k,t)S(k,t) + V_m(k,t), \tag{3}$$

where $k = 2\pi f/c$ is the wavenumber, f is the frequency, t is the time frame index, c is the speed of sound, and $P_m(k,t)$, $S(k,t)$, $H_m(k,t)$, and $V_m(k,t)$ are the STFT of $p_m(\tilde{t})$, $s(\tilde{t})$, $h_m(\tilde{t})$, and $v_m(\tilde{t})$, respectively.

In this work, our objective is to utilize a deep learning architecture for acoustic source localization in the circular harmonic domain by learning the mapping from the acquired sensor signals $P_m(k,t)$ to the DOA information using a large set of labeled training data. To this end, we propose a new learning approach where the circular harmonic representation of the observed signals is fed into a DNN framework, and the DOA estimation is formulated as an I-class classification problem, where each class corresponds to a possible angle in a set $\{\phi_1, \cdots, \phi_i, \cdots, \phi_I\}$. Then the DOA estimate is obtained as the DOA class with the highest posterior probability. Therein, the number of classes I depends on the resolution of the whole range of DOAs, and ϕ_i is the DOA corresponding to the ith class.

III. Circular Harmonic Beamforming

The aim of CHB is to combine different harmonic components to form a beam with appropriate spatial selectivity properties [37]. Ideally, the continuous circular apertures are used and the ideal beamformer response can be represented as a delta function as follows

$$B_{ideal}(k) \quad = \quad P_0\delta(\phi - \phi_s), \tag{4}$$

where P_0 is the amplitude of the impinging source and $\phi \in [-\pi, \pi)$.

It can be shown that this ideal response can be obtained by adding an infinite number of modes, so that the ideal beamformer can be written as [37], [39]

$$B_{ideal}(k) \quad = \quad \sum_{n=-\infty}^{\infty} \frac{C_n(k)}{j^n J_n(kr)} e^{jn\phi}, \tag{5}$$

where $j = \sqrt{-1}$, n is the order of harmonic. $J_n(\cdot)$ is the nth-order Bessel function of the first kind and $C_n(k)$ represents the Fourier coefficients (or circular harmonics) [21]:

$$C_n(k) \quad = \quad P_0 j^n J_n(kr\sin\eta_s) e^{-jn\phi_s}. \tag{6}$$

In real-life applications, the discretization of the continuous aperture by means of a uniform circular array with M omnidirectional sensors results in the following circular harmonics [37]

$$\tilde{C}_n(k,t) \quad = \quad \frac{1}{M}\sum_{m=1}^{M} \tilde{P}_m(k,t) e^{-jn\theta_m}, \tag{7}$$

where $\tilde{P}_m(k,t)$ is the STFT of the measured sound pressure at the mth sensor.

In (5), substituting $C_n(k)$ with $\tilde{C}_n(k,t)$ yields the nth-order circular harmonic beams response [23], [25]

$$\begin{aligned} B(k,t) &= \sum_{n=-N}^{N} \frac{\tilde{C}_n(k,t)}{j^n J_n(kr)} e^{jn\phi} \\ &= \frac{1}{M}\sum_{n=-N}^{N}\sum_{m=1}^{M} \tilde{P}_m(k,t) e^{-jn\theta_m} \frac{1}{j^n J_n(kr)} e^{jn\phi}. \end{aligned} \tag{8}$$

Note that in practice [21], the number of harmonics must be truncated to a maximum order N, which is related to the number of sensors, i.e., $N = \begin{cases} M/2-1, & M \text{ even} \\ (M-1)/2, & M \text{ odd} \end{cases}$. As a rule of thumb, $N = \lceil kr \rceil$ is usually chosen, where $\lceil \cdot \rceil$ is the ceiling function [37].

Equation (8) forms the basis of the proposed circular harmonic features discussed in the ensuing sections. To facilitate the explanation in the following sections, we re-write (8) as

$$B(k,t) \quad = \quad \sum_{n=-N}^{N} \tilde{C}_n(k,t) \cdot G_n(k) \cdot H_n(\phi), \tag{9}$$

where G_n is an equalization factor given by

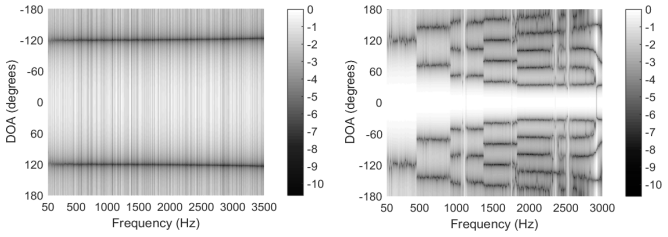
$$G_n(k) \quad = \quad \frac{1}{j^n J_n(kr)}, \tag{10}$$

and H_n is a frequency-independent phase factor,

$$H_n(\phi) \quad = \quad e^{jn\phi}. \tag{11}$$

Here, we further clarify the limitations of the conventional CHB approach. As mentioned in [25], [37], [39], we know that the DOA estimation accuracy offered by CHB is affected by the number of sensors and the radius of the array. When two factors are increased simultaneously, the performance will

be improved as this increases the maximum order that can be used. Fig. 2 shows the beampatterns with 4mic-2cm ($M = 4$ and $r = 2$ cm) and 12mic-11.9cm ($M = 12$ and $r = 11.9$ cm [39]) arrays for DOA $\phi_s = 0°$. From Fig. 2, we can notice that the beampattern of the 12mic-11.9cm array has better directivity than the 4mic-2cm array. Furthermore, the mode strengths are sensitive to reverberation and noise, which may lead to degraded DOA estimation performance for the CHB method in such environments.



(a) (b)

Fig. 2. Beampatterns for DOA $\phi_s = 0°$: (a) 4mic-2cm; (b) 12mic-11.9cm.

Thus, we aim at developing a new localization approach based on deep learning in the circular harmonic domain, which is suitable for most circular microphone arrays, especially for the small-sized array (i.e. 4mic-2cm), under variant acoustic conditions.

IV. Novel Circular Harmonic Features

Selecting suitable features is an important aspect for creating a deep model for the localization problem. In this section, we introduce three novel features based on circular harmonic.

A. Equalized Circular Harmonic (ECH) Features

Using the model presented in (7) for circular harmonics representation, the vector of circular harmonics can be written as the following structure

$$\begin{aligned}\tilde{\mathbf{C}}(k,t) &= [\tilde{C}_0(k,t), \tilde{C}_1(k,t), \tilde{C}_{-1}(k,t), \cdots, \\ &\quad \tilde{C}_N(k,t), \tilde{C}_{-N}(k,t)]. \end{aligned} \tag{12}$$

Next, we form the equalization coefficients in a vector for each frequency as follows

$$\mathbf{G}(k)=[G_0(k), G_1(k), G_{-1}(k), \cdots, G_N(k), G_{-N}(k)], \tag{13}$$

where the elements $G_n(k)$ are computed as in (10).

Thus, using (12) and (13), the equalized circular harmonic can be calculated as

$$\begin{aligned}\mathbf{C}^{ECH}(k,t) &= \tilde{\mathbf{C}}(k,t) \circ \mathbf{G}(k) \\ &= [C_0^{ECH}(k,t), C_1^{ECH}(k,t), C_{-1}^{ECH}(k,t), \\ &\quad \cdots, C_N^{ECH}(k,t), C_{-N}^{ECH}(k,t)]. \end{aligned} \tag{14}$$

where $\circ$ represents the Hadamard product operator, and the elements $C_n^{ECH}(k,t)$ are computed as

$$C_n^{ECH}(k,t) = \tilde{C}_n(k,t) \cdot G_n(k), n = -N, \cdots 0, \cdots N, \tag{15}$$

In the following, we use the real and imaginary components of the equalized circular harmonic as features. Therefore, we obtain $(2N + 1) \times 2$ components, which can be represented as,

$$\begin{aligned}\mathbf{F}^{ECH}(k,t) = \Big[&[C_0^{ECH}(k,t)]^R, [C_0^{ECH}(k,t)]^I, \ldots, [C_N^{ECH}(k,t)]^R, \\ &[C_N^{ECH}(k,t)]^I, [C_{-N}^{ECH}(k,t)]^R, [C_{-N}^{ECH}(k,t)]^I\Big]^T, \end{aligned} \tag{16}$$

where the superscripts $(\cdot)^R$ and $(\cdot)^I$ stand for real part and imaginary part of a complex number, respectively, and $(\cdot)^T$ represents the transpose of a matrix.

In addition, it can be observed that, for the ECH features, the input size of the network is $2(2N + 1) \times N_{fb}$ (number of frequency bins) in this study. In our experiments, this feature was not normalized, as we empirically observed that normalization can lead to performance degradation (results omitted due to space limitation).

B. Circular Harmonic Modes Magnitude and Phase (CH-MMP) Features

In the case of far-field DOA estimation, the phase component of the received signal at multiple microphones contributes to DOA estimation [38]. For this reason, we consider simultaneously using the magnitude and phase of the circular harmonic modes as features.

According to (11), we firstly define the following frequency-independent phase matrix $\mathbf{P_H}$, which is formed by I different weighting vectors covering the azimuth range $\phi_i \in [-\pi, \pi)$, $i = 1, \cdots, I$. Herein, I represents the number of DOA classes, which depends on the resolution of the whole range of DOAs, as aforementioned in Section II, and ϕ_i is the DOA corresponding to the ith class,

$$\mathbf{P_H} = [\mathbf{H}(\phi_1), \cdots, \mathbf{H}(\phi_i), \cdots, \mathbf{H}(\phi_I)], \tag{17}$$

where

$$\mathbf{H}(\phi_i) = \left[e^{j0\phi_i}, e^{j1\phi_i}, e^{j(-1)\phi_i}, \cdots, e^{jN\phi_i}, e^{j(-N)\phi_i}\right]^T. \tag{18}$$

Thus, using (14) and (18), the magnitude and phase matrix of the circular harmonic modes can be calculated as

$$\begin{aligned}\mathbf{C}^{MMP}(k,t,\phi) &= \mathbf{C}^{ECH}(k,t) \times \mathbf{P_H} \\ &= [C^{MMP}(k,t,\phi_1), \cdots, C^{MMP}(k,t,\phi_i), \\ &\quad \cdots, C^{MMP}(k,t,\phi_I)], \end{aligned} \tag{19}$$

where

$$\begin{aligned}C^{MMP}(k,t,\phi_i) &= \mathbf{C}^{ECH}(k,t) \cdot \mathbf{H}(\phi_i) \\ &= \sum_{n=-N}^{N} \tilde{C}_n(k,t) \cdot G_n(k) \cdot H_n(\phi_i). \end{aligned} \tag{20}$$

Therefore, the feature of the circular harmonic modes magnitude and phase, which contains I components, can be expressed as

$$\mathbf{F}^{MMP}(k,t,\phi) = \left[\mathbf{C}^{MMP}(k,t,\phi)\right]^T, \tag{21}$$

In order to make an equal contribution to each TF unit, we normalize the elements of $\mathbf{F}^{MMP}(k,t,\phi)$. Since the power spectrum peak contains information about DOA, we use the power spectrum as input. Note that, the difference between the CH-MMP and the ECH features is that the latter does not include the term $H_n(\phi)$ in (9).

In addition, it can be observed that, for the CH-MMP features, the input size of the network is $I \times N_{fb}$ in this study.

C. Circular Harmonic Enhanced Modes Magnitude and Phase (CH-E-MMP) Features

Since mode strengths are sensitive to noise and reverberation, this may degrade the performance of the two features aforementioned for the DOA estimation in adverse environments. To address this issue, we propose another feature based on circular harmonic enhanced modes magnitude and phase, to further improve their robustness against noise and room reverberation.

In our earlier work [26], we demonstrated that DOA estimation accuracy can be improved by selecting TF bins of higher power, which is often an indication for an active source at this direction [40]. Furthermore, as mentioned in [38], with an increase in n, the mode strength is decreasing. As a result, the mode strengths at higher orders become small, and are thus prone to the corruption by noise and reverberation. For these reasons, we employ the power of $n = 0$ mode strength, which represents omnidirectional fields that have no variation in the azimuth direction, when compared with mode strength of other orders. This can help us more accurately find the useful TF bins with high power in the circular harmonic domain.

The power of the 0-th mode strength, which has no variation in the azimuth direction, on the basis of TF units can be calculated by the following equation

$$E_0(k,t) = \left| \tilde{C}_0(k,t) \cdot G_0(k) \cdot H_0(\phi_i) \right|^2 . \tag{22}$$

Subsequently, we sort the powers according to their levels

$$E_0(k_1,t_1) \geq E_0(k_2,t_2) \cdots \geq E_0(k_q,t_q) \cdots , \tag{23}$$

where $E_0(k_1,t_1)$ corresponds to the highest power at (k_1,t_1) TF bin, $E_0(k_2,t_2)$ corresponds to the second highest, and $q = 1,2,\cdots,Q$ with Q representing the total number of TF bins.

The next step is to use the sorted powers to form the enhanced function $E_f(k,t)$, as follows

$$E_f(k,t) = \begin{cases} 1, & \text{if } E_0(k_q,t_q) \geq E_0(k_u,t_u) \\ 0, & \text{otherwise,} \end{cases} \tag{24}$$

where $E_0(k_u,t_u)$ is the power at the uth TF bin that $u = \alpha Q$, with $\alpha \in (0,1]$ being a pre-defined threshold, whose selection will be discussed in Section V-C. The reason for choosing the binary masks instead of soft masks is that using binary masks, we can obtain a cleaner signal by selecting the reliable TF bins dominated by the target sound and removing those from noise and reverberation. However, if the soft mask is used, some interference may still remain in the signal, which can lead to inaccurate DOA estimation.

Therefore, the feature of circular harmonic enhanced modes magnitude and phase, which contains I components, can be represented as

$$\begin{aligned} \mathbf{F}^{E-MMP}(k,t,\phi) \quad = \quad & \Big[C^{E-MMP}(k,t,\phi_1), \cdots , \\ & C^{E-MMP}(k,t,\phi_i), \cdots , C^{E-MMP}(k,t,\phi_I) \Big]^T , \end{aligned} \tag{25}$$

where $C^{E-MMP}(k,t,\phi_i)$ can be obtained by substituting (24) into (20) as follows

$$C^{E-MMP}(k,t,\phi_i) = \sum_{n=-N}^{N} E_f(k,t) \cdot \tilde{C}_n(k,t) \cdot G_n(k) \cdot H_n(\phi_i), \tag{26}$$

Similar to the CH-MMP features, we normalize the elements of $\mathbf{F}^{E-MMP}(k,t,\phi)$ and use the power spectrum as input, which has size $I \times N_{fb}$. Note that, different from the ECH features, we have applied normalization to the CH-MMP and CH-E-MMP features to make the beam pattern in these features look smoother, as often done in the literature of circular harmonic beamforming. However, we empirically observed that the performance difference between the normalized and un-normalized version for these two features is negligible.

V. Practical Systems for DOA Estimation

After obtaining the input features, we can learn a mapping from the input features to the DOA classes. In this section, we develop a CNN framework for DOA estimation of an acoustic source with these features as inputs.

A. CNN Network Architecture

Many different network architectures could be employed for the DOA classification purpose. CNNs have a property, namely, translation shift-invariant [42], which means that the shifts in input would lead to shifts in the output, otherwise, they would remain unchanged. With this property, if the CNN has learned a circular harmonic feature useful for detecting a DOA angle during training, it is expected to capture the similar feature related to such a DOA during testing [43]. Therefore, we choose the popular CNNs, as they are translation shift-invariant, and also robust to unseen acoustic scenarios [41]. The detailed CNN that we used is composed of an input layer, few convolutional layers, two fully connected layers, and an output layer. The activation function of convolutional layers and fully connected (FC) layers is rectified linear units (ReLU). Dropout with a rate 0.5 is used between the convolutional layer and the FC layer, and after each FC layer, which is used to mitigate overfitting. Each convolutional layer has 64 local filters of size 3×3 to learn local correlation at local frequency regions. Since resolution is important for the accurate estimation of an acoustic source, max pooling (or any other down-sampling after each convolutional layer) is not used in our work. The architecture of the proposed CNN is illustrated in Fig. 3.

In this study, the three DOA estimation approaches based on deep learning in the circular harmonic domain are ECH-CNN, CH-MMP-CNN and CH-E-MMP-CNN, respectively. We should note that, with a more sophisticated network [44], the ECH feature may be exploited more efficiently. However, the focus of this work is on the performance of the features designed with circular harmonics, therefore, we have chosen CNN as the network architecture, despite the availability of a wide range of network architectures including end-to-end systems.

The given circular harmonic features $\mathbf{F}$, namely $\mathbf{F}^{ECH}$, $\mathbf{F}^{MMP}$ and $\mathbf{F}^{E-MMP}$, which contain information about DOA, are first input to the convolutional layer [45], and the corresponding outputs are determined according to

$$\begin{aligned} \mathbf{X} \quad = \quad & f(\mathbf{W}_{c4} * f(\mathbf{W}_{c3} * f(\mathbf{W}_{c2} * f(\mathbf{W}_{c1} * \mathbf{F} + \\ & \mathbf{b}_{c1}) + \mathbf{b}_{c2}) + \mathbf{b}_{c3}) + \mathbf{b}_{c4}), \end{aligned} \tag{27}$$

where $\mathbf{W}_{c1}$, $\mathbf{W}_{c2}$, $\mathbf{W}_{c3}$ and $\mathbf{W}_{c4}$ refer to the weight of the convolution kernel corresponding to each convolutional layer, respectively, and $\mathbf{b}_{c1}$, $\mathbf{b}_{c2}$, $\mathbf{b}_{c3}$ and $\mathbf{b}_{c4}$ represent an additive

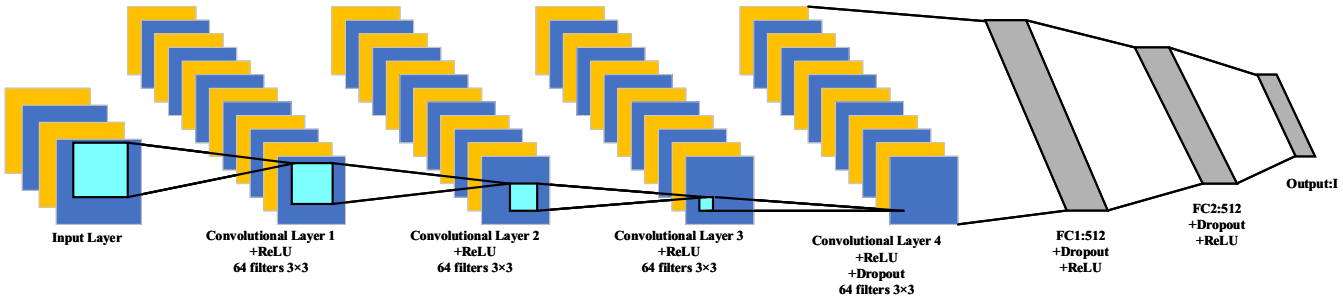


Fig. 3. The architecture of the proposed CNN for acoustic source localization. Each convolutional layer has 64 filters and is followed by a ReLU layer. After the convolutional layers, a fully connected (FC) layer is used (followed by ReLU activation functions). Dropout layers are employed between the convolutional layer and the FC layer and after each FC layer, to mitigate potential overfitting problem.

bias corresponding to each convolutional layer, respectively. The activation function f is chosen as ReLU [45], which is defined as $f(x) = \max(0, x)$.

The cross-entropy [45] is used as the loss function, which is given by

$$L = -\sum_{i=1}^{I} g_i \log(P_{CNN}(\phi_i|\mathbf{F})), \tag{28}$$

where g_i denotes the ground-truth label corresponding to the ith class, and $P_{CNN}(\phi_i|\mathbf{F})$ stands for the posterior probability of the DOA candidates for the acoustic source, which can be written as

$$P_{CNN}(\phi_i|\mathbf{F}) = \frac{exp(o_i)}{\sum_{i=1}^{I} exp(o_i)}, \tag{29}$$

where o_i is the output value of the output layer corresponding to the ith class.

In the final layer, we use the softmax activation function [45] to perform classification. The softmax function generates the posterior probability for each of the ith class. The final source DOA is estimated by maximizing the posterior probability, i.e.,

$$\hat{\phi}_s = \arg\max_{\phi_i}(P_{CNN}(\phi_i|\mathbf{F})), \tag{30}$$

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

We use stochastic gradient descent with momentum (SGDM) [45] as the optimizer. The value of minibatches is set as 64, 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.

B. The Block Diagram of the Proposed Algorithm

Fig. 4 illustrates the entire process of the proposed method, which consists of a training and a test phase. In the training phase, the CNN is trained with a data set that consists of feature vectors of fixed dimension based on circular harmonics, namely CH-MM, CH-MMP and CH-E-MMP, and the corresponding true DOA class labels. This stage corresponds to the blue boxes in Fig. 4. In the test phase, given microphone signals, our aim is to estimate the posterior probability of each DOA class based on the input feature representations on the basis of circular harmonic derived from the microphone signals. Finally, the DOA estimate is obtained by selecting the DOA class with the highest probability. This stage corresponds to the green boxes in Fig. 4.

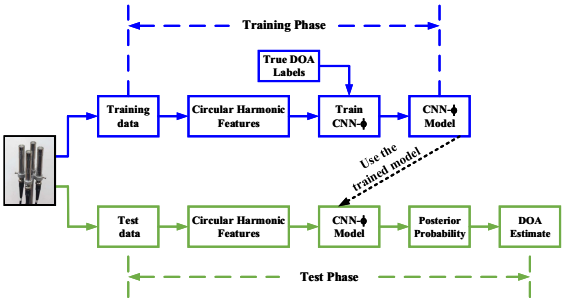


Fig. 4. Block diagram illustrating the proposed DOA estimation method.

VI. Experimental Evaluations

This section studies the performance of the proposed method through simulations and real data experiments. DOA estimation using the proposed method is investigated and compared to baseline methods. Experiments 1 and 2 are designed to evaluate the performance of the proposed algorithm in the presence of room reverberation and background noise at different levels, respectively, which are tested in two situations: fixed-room and changeable-room. Experiments 3, 4 and 5 investigate the influence of array sizes, microphone imperfections and elevations, respectively. Lastly, the algorithms are further tested with real data. The section starts with a description of datasets, evaluation metrics and parameter setup, and then presents the experimental results.

A. Datasets

The simulated data used for training were generated in a fixed-room and changeable-room, respectively, and the detailed configurations are shown in Table I, Table II and Fig. 5. Since our proposed algorithm is suitable for most circular arrays, we chose to use the small-sized array (namely M = 4 and r = 0.02 m) in our experiments, which can demonstrate the flexibility of our proposed method.

In Fig. 5, the UCA with M = 4 equidistant omnidirectional sensors and the diameter d = 0.04 m, was placed in the center of the room, coinciding with the origin of the x and y axes. The speaker was located at the same height as the microphone array, namely $\eta_s = 90°$, with distance from the speaker to the center of the array being 1.5 m. Subsequently, we considered that the spatial resolution for source location was set to be $10°$ in the simulations. Thus, the total number of candidate source locations was 36, namely $I = 36$, which were distributed

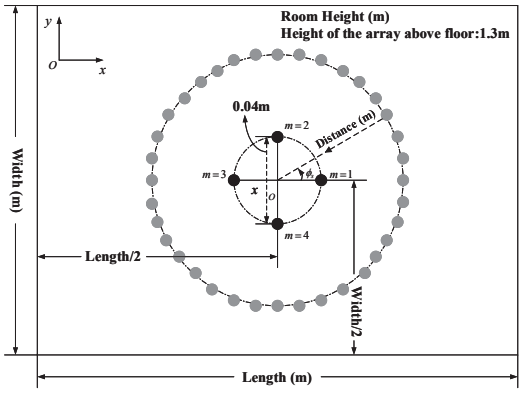


Fig. 5. Illustration of the simulation setup. The black solid dots, distributed uniformly on the dash-dotted circle with a diameter of 0.04 m, denote the UCA. The gray circles, distributed around the dash-dotted circle, denote the acoustic sources.

uniformly from $-180°$ to $170°$, shown as the dash-dotted circle in Fig. 5. Since we have 36 classes, the sources whose DOAs deviate from these discretized DOAs, i.e. off the grid, will be assigned with one of these 36 classes. For example, the DOA of a source at $173°$ would be estimated as $170°$. To generate the RIRs [46] from acoustic sources to sensors, we used a software that was based on the well-known image method for simulating a reverberant environment in a room [47].

Three types of noise (i.e., Babble and Destroyerops which are from the Noisex-92 dataset [48], and white Gaussian noise (WGN)) were used as the background noise sources. Several levels of room reverberation with various reverberation times were tested, which will be specified later. The sound speed was 340 m/s. Speech signals of 0.5 s length, sampled at 16 kHz, were chosen randomly from the well-known TIMIT speech database [49]. The utterances in TIMIT are continuous speech. For intermittent speech, there are silent periods, in which case, the proposed method would need to be slightly modified to include a class of *no direction* as in [50]. More specifically, we would have $I = 37$ that corresponds to 36 directions and 1 *no direction*. The source was convolved with the simulated RIRs from the source to every microphone. For all the evaluated algorithms, the STFT was calculated using a Hamming window of 1024 samples with 50% overlap between consecutive frames and the number of frequency bins was 511, namely $N_{fb} = 511$. Thus, for the CH-MMP and CH-E-MMP features, the input size of CNN is 36×511. While for the ECH features, the input size of CNN is $2(2N+1) \times 511$, and its exact size is determined by the value of maximum order N. Specifically, when $N = 1$, the input size is 6×511, as a result, we can use at most two convolutional layers. Except for this special case, we all use four convolutional layers in our designed CNN network.

In the following, to analyze the source localization performance systematically, we considered six different aspects in the simulations. For each experiment, the data used for training, validating and testing are described below:

(1) *Effect of Room Reverberation*: To evaluate the influence of reverberation on the performance of the proposed method, we considered two different scenarios, i.e., a fixed-room and changeable-room, under varying levels of reverberation times in the first experiment. As aforementioned, the RT_{60} was varied from 200 to 800 ms with a step increase of 100 ms, the SNR is 10 dB and the number of candidate source locations is 36. Under the fixed-room condition, 250 utterances from TIMIT database were randomly selected, which outputs $250 \times 36 \times 7 = 63000$ training signals, while another 60 utterances

TABLE I
FIXED-ROOM CONFIGURATIONS IN TRAINING PROCESS AND TEST PROCESS

Reverberant Conditions				
Type	Length(m) × Width(m) × Height(m)	Distance(m)	RT_{60}(ms)	SNR(dB)
Train	$9.7 \times 7.05 \times 3.0$	1.5	picked randomly between 200 and 800	10
Test	$9.7 \times 7.05 \times 3.0$	1.5	200 to 800 with an increment of 100	10
Noisy Conditions				
Type	Length(m) × Width(m) × Height(m)	Distance(m)	SNR(dB)	RT_{60}(ms)
Train	$9.7 \times 7.05 \times 3.0$	1.5	picked randomly between -5 and 20	300
Test	$9.7 \times 7.05 \times 3.0$	1.5	-5 to 20 with an increment of 5	300

TABLE II
CHANGEABLE-ROOM CONFIGURATIONS IN TRAINING PROCESS AND TEST PROCESS

Reverberant Conditions					
Type	Room	Length(m) × Width(m) × Height(m)	Distance(m)	RT_{60}(ms)	SNR(dB)
Train	1 2 3	$6.0 \times 3.0 \times 3.0$ $8.0 \times 5.5 \times 3.0$ $14.0 \times 11.0 \times 3.0$	1 1.5 2	picked randomly between 200 and 800	10
Test	–	$11.0 \times 8.5 \times 3.0$	1.75	200 to 800 with an increment of 100	10
Noisy Conditions					
Type	Room	Length(m) × Width(m) × Height(m)	Distance(m)	SNR(dB)	RT_{60}(ms)
Train	1 2 3	$6.0 \times 3.0 \times 3.0$ $8.0 \times 5.5 \times 3.0$ $14.0 \times 11.0 \times 3.0$	1 1.5 2	picked randomly between -5 and 20	300
Test	–	$11.0 \times 8.5 \times 3.0$	1.75	-5 to 20 with an increment of 5	300

were utilized as the validation data, which output $60 \times 36 \times 7 = 15120$ signals in the validation set. In the testing phase, for each candidate source location, we randomly selected another 20 utterances to generate a test set of $20 \times 36 \times 7 = 5040$ signals. Then, in the changeable-room condition, the acoustic localization algorithms were trained in the three different rooms (as mentioned in Table II) and tested in one room. As a result, the training data was $250 \times 36 \times 7 \times 3 = 189000$ signals and the validation data was $60 \times 36 \times 7 \times 3 = 45360$ signals, while the testing set had $20 \times 36 \times 7 = 5040$ signals.

(2) *Effect of Noise Level*: The second experiment was carried out to investigate the influence of noise on the proposed method. In this experiment, all the configurations were the same as the first experiment, except the SNR which is varied from -5 to 20 dB with a step increase of 5 dB with the reverberation time of $RT_{60} = 300$ ms. Thus, for the fixed-room condition, there were $250 \times 36 \times 6 = 54000$, $60 \times 36 \times 6 = 12960$ and $20 \times 36 \times 6 = 4320$ signals for training, validating and testing, respectively. Similarly, for the changeable-room condition, the training set was composed of $250 \times 36 \times 6 \times 3 = 162000$ signals and the validation set had $60 \times 36 \times 6 \times 3 = 38880$ signals, while the testing set had $20 \times 36 \times 6 = 4320$ signals.

(3) *Effect of Array Size*: This experiment was designed to evaluate the performance of array size. The data were simulated in the rectangular room, whose dimensions are 9.7 m $\times$ 7.05 m $\times$ 3 m, with a reverberation time of 300 ms and a noise level of 10 dB. The array size d was varied from 0.02 m to 0.3 m (i.e., 0.02 m to 0.1 m with a step increase of 0.01 m and 0.12 m to 0.3 m with a step increase of 0.02 m). Thus, we had 19 cases of array size. As a result, there were $250 \times 36 \times 19 = 171000$ signals for training, $60 \times 36 \times 19 = 41040$ for validating and $20 \times 36 \times 19 = 13680$ for testing.

(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 experiment, we compare the performance of the various methods in the presence of microphone gain and phase errors. The dimensions of the simulated rectangular room are 9.7 m $\times$ 7.05 m $\times$ 3 m, with a reverberation time of 300 ms and a noise level of 10 dB. Herein, assume that the microphone gain and phase errors are unknown and bounded, respectively, by

$$\varepsilon \in \lambda[-0.1, 0.1], \psi \in \lambda[-5^\circ, 5^\circ], \tag{31}$$

where ε represents the microphone gain error, ψ represents the microphone phase error, and λ is a scale parameter used to control the error ranges, which is varied from 0 to 1 with a step increase of 0.2, i.e., with the microphone mismatches being gradually increased. In the simulations, the gain and phase errors of each microphone were randomly selected within the error range given by (31). In total, the training set had $250 \times 36 \times 6 = 54000$ signals, the validation set had $60 \times 36 \times 6 = 12960$ signals and the testing set had $20 \times 36 \times 6 = 4320$ signals.

(5) *Effect of Elevation*: This experiment aimed at assessing the impact of elevation angles on our proposed three methods. We considered different elevation angles, i.e., $\eta_s = 30^\circ$, 60° and 90°, in the simulated rectangular room of 9.7 m $\times$ 7.05 m $\times$ 3 m, respectively. The other datasets were the same as described in Table I.

(6) *Effect of Temperature*: We include an analysis for the potential impact of the indoor temperature on the performance of the proposed method.

To further evaluate the effectiveness of the proposed method, we also selected 20 utterances and recorded them in a real rectangular conference room with dimensions of approximately 9.7 m $\times$ 7.05 m $\times$ 3 m and a reverberation time of 350 ms. A small-sized array was placed horizontally around the center of the room, and the other conditions resembled those in the above simulations. A photograph of the real experiment and microphone array is shown in Fig. 6. In addition, we used the publicly available LOCATA dataset [51], [52] to evaluate our proposed method.

The sensors used in the real-world experiments were all 1/2-inch sensors (MPA201; BSWA Technology Co., Ltd.). The received sensor signals were sampled at 16 kHz through a data-acquisition device (NI-USB-4432 and cDAQ-9178; National Instruments) with 24-bit. In the real experiments, the actual speaker locations were determined using protractors and rulers, and the spatial resolution is 20° (i.e., -170° to 170° with 18 candidate positions, namely $I = 18$). As a result, there were $20 \times 18 = 360$ signals for testing.

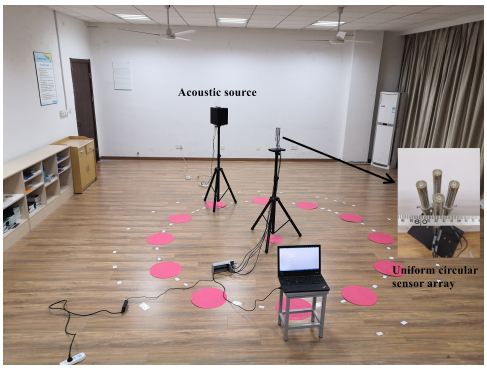


Fig. 6. Photography showing the real-world experiment and the uniform circular sensor array used. The diameter of the circular microphone array is 0.04 m.

B. Evaluation Metrics

To facilitate evaluations, we use the localization accuracy (Acc) as performance metrics, which is defined as:

$$Acc = \frac{N_{cr}}{N_e} \times 100\%, \tag{32}$$

where N_e represents the number of source locations being evaluated, and N_{cr} denotes the number of source locations that are correctly recognized. Herein, an acoustic source is considered being correctly localized if the deviation of the estimated DOA from the actual DOA is within $\pm\phi_0$ for a spatial resolution of ϕ_0, namely, $\pm 10^\circ$ for the simulations and $\pm 20^\circ$ for the real data experiments.

C. Parameter Setup

For the pre-defined threshold α discussed in Section IV-A, we tested $\alpha \in [0.1, 0.5]$ for finding the suitable uth TF bin when the total number of TF bins $Q = 14322$. Since the proposed method is data-driven, we empirically choose $\alpha = 0.2$, which seems to be appropriate for most scenarios.

An example is given in Fig. 7, where we plot the normalized histograms of original TF bins (grey bins) and suitable TF bins (blue bins) for DOA $\phi_s = 60^\circ$ with (a) SNR $= 0$ dB, $RT_{60} = 300$ ms and (b) SNR $= 10$ dB, $RT_{60} = 700$ ms, respectively. From Fig. 7, it is noticed that a certain amount of suitable TF bins were calculated by using $\alpha = 0.2$ and demonstrated in the normalized histograms in adverse environments (e.g. low SNR or high reverberation), which further confirms the validity of the enhanced function in the circular harmonic domain. We have also evaluated the DOA estimation accuracy by using the CH-E-MMP-CNN method for the fixed room under reverberant and noisy conditions, for $\alpha = 0.2$ and 0.6. We found that $\alpha = 0.2$ offers better results as compared with $\alpha = 0.6$. Such results were not shown here due to the space limitation.

D. Baseline Methods

The performance of the proposed ECH-CNN, CH-MMP-CNN and CH-E-MMP-CNN are evaluated and compared with several baselines including the STFT-CNN [29], [30], GCC-PHAT-CNN [9], [27], IRM-BLSTM [33] and PSM-BLSTM [12] methods in both simulated and real room environments. The environmental conditions and noise types are the same as those aforementioned.

STFT-CNN: This is a broadband DOA estimation method based on CNN, in which phase component of the short-time Fourier transform (STFT) coefficients of the received microphone signals are used as features and directly fed into the CNN. The STFT-CNN model consists of three convolutional layers, each containing 64 small filters of size 2×2 and two fully connected layers with each having 512 units. At the end of the three convolution layers and after each fully connected layer, dropout with a rate of 0.5 was used to mitigate the potential overfitting problem. The activation function used in the output layer is softmax and the others are ReLU. Cross-entropy is used as the loss function and Adam is used as the optimizer.

GCC-PHAT-CNN: In this method, the popularly used features, i.e., generalized cross correlation with phase transform (GCC-PHAT), are extracted from pairs of microphone signals and provided as input to the CNN framework for sound source localization. The GCC-PHAT-CNN model consists of three convolutional layers, each containing 64 small filters of size 3×3, two max pooling layers and one fully connected layer. A batch normalization layer is used after each convolutional layer. The activation function of the output layer is softmax and those of the others are ReLU. Cross-entropy is used as the loss function and Adam is used as the optimizer.

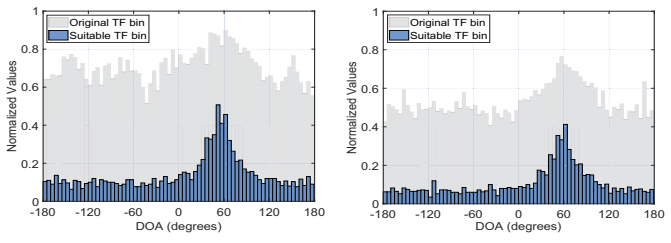


Fig. 7. Normalized histograms of original TF bins (grey bins) and suitable TF bins (blue bins) for DOA $\phi_s = 60^\circ$ with choosing $\alpha = 0.2$: (a) SNR $= 0$ dB and $RT_{60} = 300$ ms; (b) SNR $= 10$ dB and $RT_{60} = 700$ ms.

IRM-BLSTM and *PSM-BLSTM*: Both methods are guided by the TF masks. The ideal ratio mask (IRM) is ideal for speech enhancement only when the mixture phase is the same as the clean phase at each TF unit. While the phase-sensitive mask (PSM) takes the phase difference into consideration by scaling down the ideal mask when the mixture phase is different from the clean phase using a cosine operation. The bi-directional long short-term memory (BLSTM) is trained to estimate the IRM and PSM, respectively, yielding better mask estimation for localization. The BLSTM contains two hidden layers each with 600 units in each direction. Sigmoidal units are used in the output layer. The Adam algorithm is used to minimize the mean squared error for mask estimation.

E. Source Localization Results in Simulation Experiments

1) Effect of Room Reverberation: In our first set of simulations, we investigated the localization accuracy under different room reverberation. Fig. 8 and Fig. 9 show the localization accuracy of each method under the conditions of the fixed-room and changeable-room, respectively, when the reverberation time RT_{60} is varied from 200 to 800 ms with a step increase of 100 ms and the level of noise in terms of SNR is 10 dB, in Babble, Destroyerops and WGN noise types. In general, the performance of all tested algorithms degrade with the increase in the level of reverberation. The proposed CH-E-MMP-CNN outperforms all the methods including CH-MMP-CNN. With the circular harmonic enhanced function, more reliable TF bins, which are dominated by the sound source, are selected, leading to improved localization performance, especially for the changeable-room in WGN noise. In contrast, the ECH-CNN method performs less effectively, which could be ascribed to the use of the equalized circular harmonic, thus lacking the related phase information, which is crucial for far-field acoustic localization when the microphones are placed close to each other. The IRM-BLSTM and PSM-BLSTM methods perform well except for the Destroyerops noise, in which persistent interference appears at low frequencies, caused by the background noise from the operating room in the destroyer. In addition, according to [33], the IRM/PSM masks are more accurately estimated at speech onsets and lower frequencies, probably because the energy of the direct speech is relatively stronger in these TF regions. These probably are the reasons why these two mask-based methods are sensitive to Destroyerops noise. As a result, both methods degrade by around 30% in reverberation.

The STFT-CNN and GCC-PHAT-CNN methods show different results for different room environments. In the fixed-room condition, the accuracy of both methods is mostly over 95% for Babble and Destroyerops noise, whereas the performance degrades for WGN, with an accuracy just over 60% for all the reverberation levels. In the changeable-room condition, GCC-PHAT-CNN performs well in Babble and Destroyerops noises, with an accuracy at approximately 95%, but degrades 25% in WGN noise. The performance of STFT-CNN decreases with the increase in the reverberation level in all three types of noise. The reason that GCC-PHAT-CNN performs better than STFT-CNN is that the PHAT weighting function used in GCC-PHAT-CNN can reduce the degradation from reverberation, while the spectrum used in STFT-CNN is prone to the degradation by room reverberation.

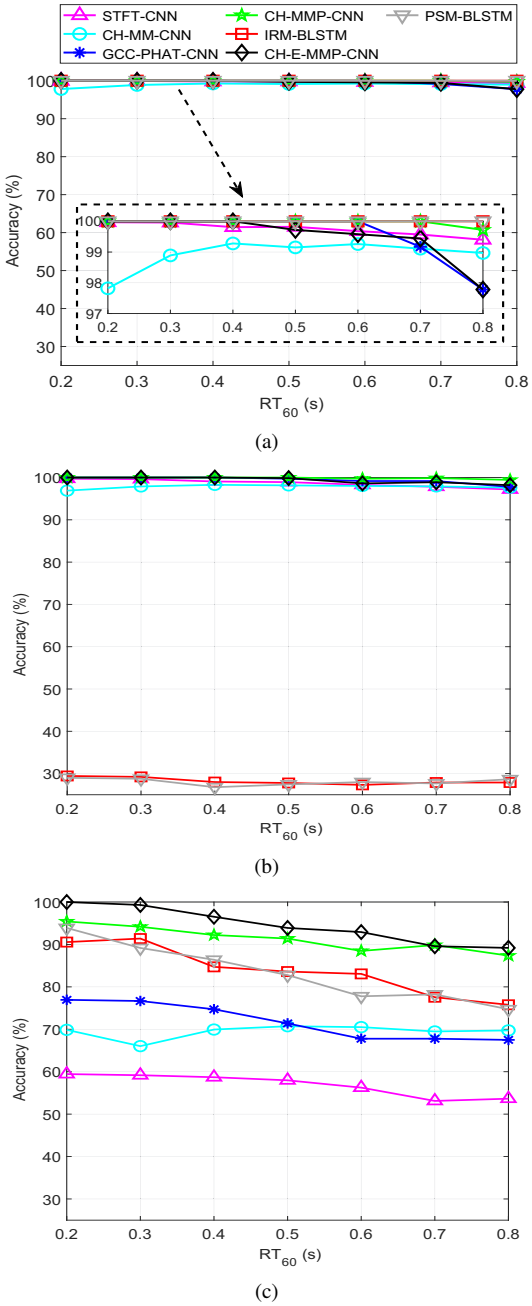


Fig. 8. Effect of room reverberation on the performance of each method for localization accuracy in the fixed-room with SNR = 10 dB: (a) Babble; (b) Destroyerops; (c) WGN.

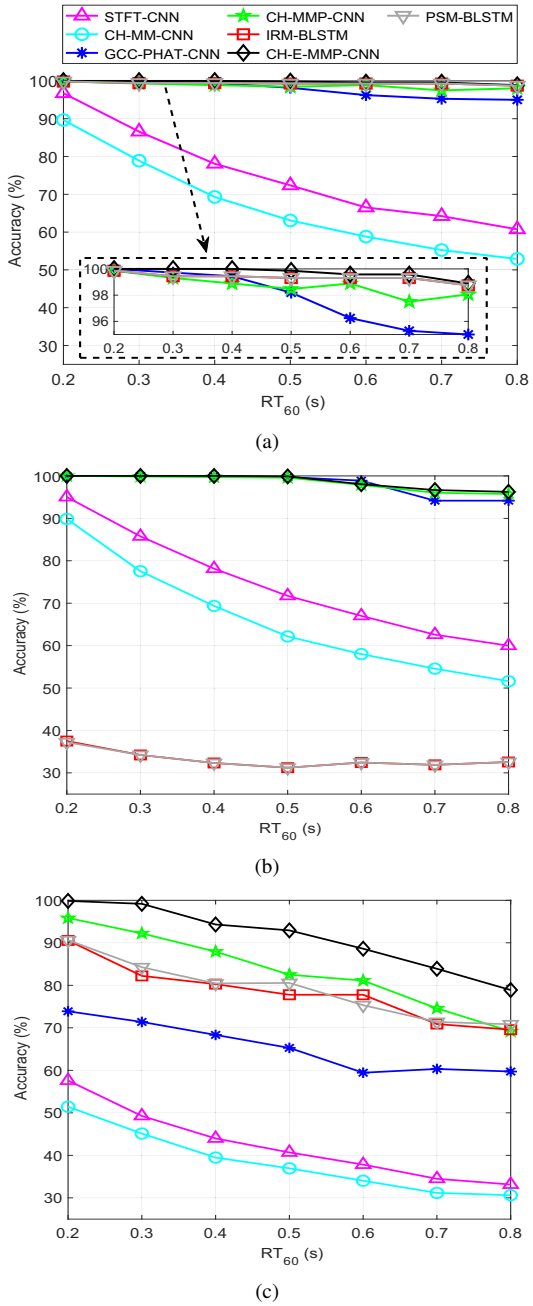


Fig. 9. Effect of room reverberation on the performance of each method for localization accuracy in the changeable-room with SNR = 10 dB: (a) Babble; (b) Destroyerops; (c) WGN.

2) Effect of Noise Level: Fig. 10 and Fig. 11 show the localization accuracy of each method under different noise types (i.e., Babble, Destroyerops and WGN) under the fixed-room and changeable-room conditions, respectively, when the SNR is varied from -5 to 20 dB with a step increase of 5 dB with the reverberation time of $RT_{60} = 300$ ms. The proposed CH-E-MM-CNN method offers better localization performance under all three noise types in different room conditions as compared with other methods including the proposed CH-MMP-CNN method, which degrades at low SNRs, i.e., SNR = 0 dB and -5dB. This is because the CH-E-MMP feature selects the TF bins, which are less affected by noise, to achieve more accurate localization. The proposed ECH-CNN method, however, is outperformed by other methods in all situations (except the IRM-BLSTM and PSM-BLSTM algorithms in Destroyerops noise). This could be because the ECH features lack dominant phase information and are also corrupted in low SNRs, especially in the case of WGN noise. The IRM-BLSTM and PSM-BLSTM approaches perform well in Babble noise, but they appear to be sensitive to Destroyerops noise. This is probably due to the strong energy in the low frequencies of the Destroyerops noise, which can distort the target sound significantly, as analysed in the section of Effect of Room Reverberation. The other algorithms, such as STFT-CNN and GCC-PHAT-CNN, provide good results in Babble and Destroyerops noise under fixed-room condition, whose

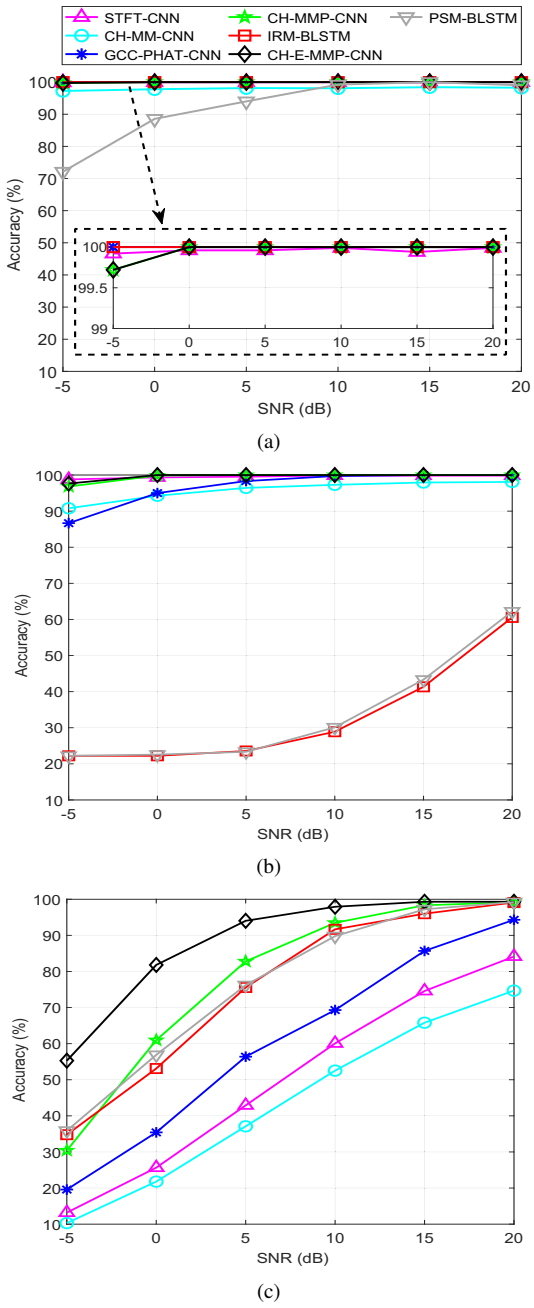


Fig. 10. Effect of noise level on the performance of each method for localization accuracy in the fixed-room with RT_{60} = 300 ms: (a) Babble; (b) Destroyerops; (c) WGN.

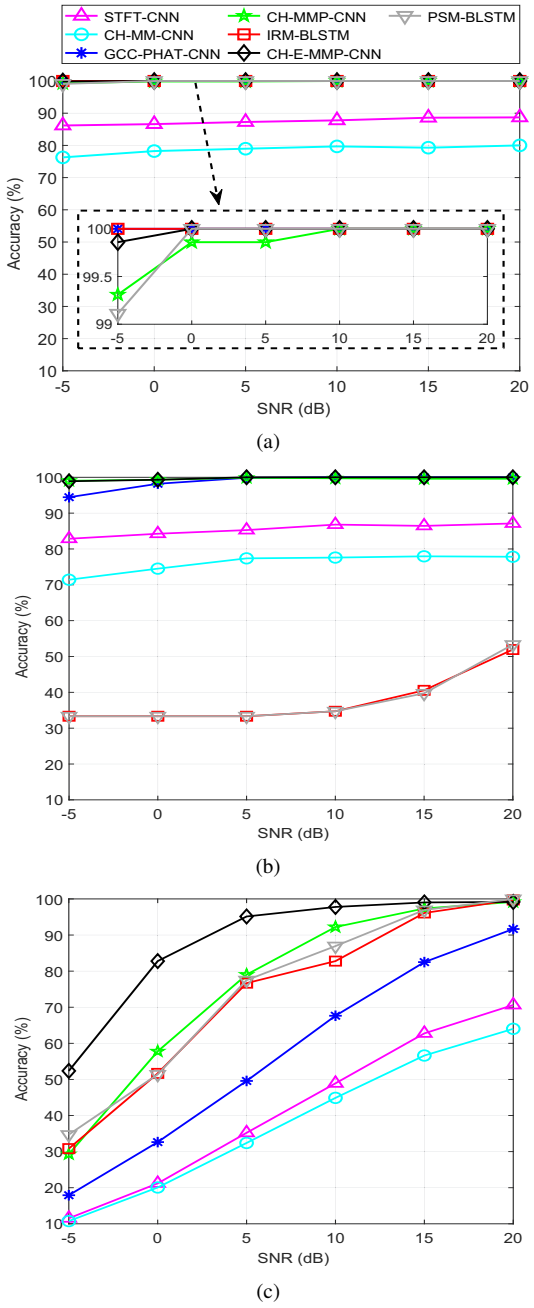
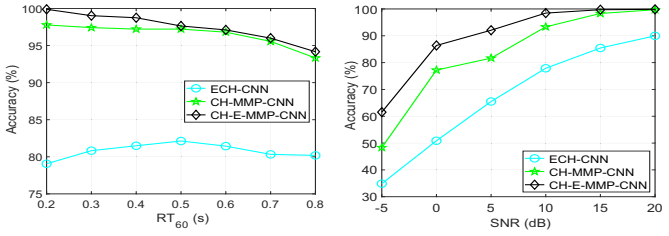


Fig. 11. Effect of noise level on the performance of each method for localization accuracy in the changeable-room with RT_{60} = 300 ms: (a) Babble; (b) Destroyerops; (c) WGN.

accuracy is nearly 95%, but under changeable-room condition, the STFT-CNN degrades to 85%, which implies that the difference between the training and test conditions may affect the DOA estimation accuracy. In addition, for WGN noise, the accuracy of both two methods degrades rapidly with the increase in noise level, and especially at low SNRs, the results are down to nearly 20%. This suggests that using the PHAT weighting function, we can potentially mitigate the detrimental impact of noise on the features learned, e.g. SNR = 0 dB and -5 dB.

Note that, for the small-sized array with M = 4 and r = 0.02 m, we have the order $N = 1$ in terms of the discussions after equation (8). Therefore, for the ECH feature, we have used two convolutional layers, which is different from CH-MMP-CNN and CH-E-MMP-CNN. To make a fair comparison, we have also performed experiments for a different array with M = 8 and r = 0.1 m, in which case, we have the order $N = 3$. This allows us to use the same number of convolutional layers, i.e. four layers, for all the three proposed methods. Fig. 12 shows the localization accuracy of the proposed methods under different room reverberation and WGN noise, respectively. From this figure, we can see that the performance of all proposed methods for the array with M = 8 and r = 0.1 m has improved over the array with M = 4 and r = 0.02 m. This is because the available order of the circular harmonic has been increased. Specifically, for

the ECH method, the dimension of the input to the network is expanded to 14 × 511. For the CH-MMP and CH-E-MMP features, they can use higher orders of circular harmonic modes information which can provide better directivity, hence their performance has also been improved.



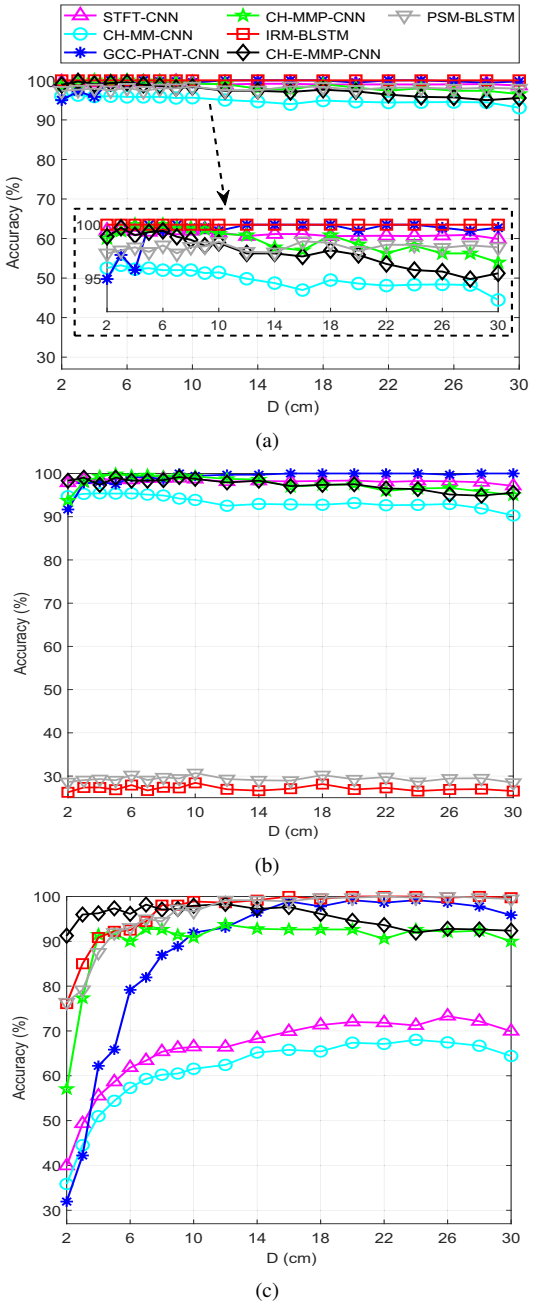
(a) (b)

Fig. 12. The localization accuracy of the proposed methods in the fixed-room: (a) room reverberation; (b) WGN noise.

We have also tested another network architecture, i.e. the combined CNN and long short term memory (LSTM) [50] network, to evaluate our proposed ECH, CH-MMP and CH-E-MMP features for different levels of noise, in the fixed-room with $RT_{60} = 300$ ms. Here, we did not perform detailed parameter tuning for this network. The CNN and LSTM network consists of two convolutional layers, one LSTM layer and a FC layer. The activation function used in the two convolutional layers and the FC layer is ReLU. Softmax is used as the activation function for the final layer. Cross-entropy is used as the loss function and the optimizer is SGDM. Dropout with a rate of 0.5 is used for the LSTM and FC layer to mitigate the potential overfitting problem. Each convolutional layer has 16 local filters of size 3×3 and the number of nodes for LSTM and FC layer is 600 and 1024, respectively. There are no subsampling layers after the convolutional layers. All three proposed methods still work in this architecture, which can demonstrate the flexibility of our proposed features. The CH-E-MMP feature performs the best among three, giving an accuracy of 88.21%.

3) Effect of Array Size: Fig. 13 shows the localization accuracy of each method when the array diameter d is varied from 0.02 to 0.3 m with $RT_{60} = 300$ ms and SNR = 10 dB. We can see that the performance of the compared approaches has been improved with the increase in array size. With the proposed CH-E-MMP-CNN method, more reliable TF bins are selected due to the use of an enhanced function with circular harmonics which results in the accuracy greater than 90% in different types of noise. Using a small scale array, e.g. d = 0.02 and 0.03 m, the performance of the proposed CH-MMP-CNN degrades to lower than 80%. This implies that the array size impacts significantly on the DOA estimation accuracy. In contrast, due to the lack of phase information in the features, the ECH-CNN method still gives the worst result, except in the case for the Destroyerops noise. The remaining algorithms, i.e. the IRM-BLSTM, PSM-BLSTM and STFT-CNN, all perform well in Babble noise, but degrade significantly in other two types of noise. To be specific, for different size array, the IRM-BLSTM and PSM-BLSTM show better performance in WGN but worse performance in Destroyerops noise, which illustrate the unstable performance of these two methods in different types of noise, for similar reasons we already discussed in the previous sections. The STFT-CNN performs well in Destroyerops noise but poorly in WGN noise (i.e., the accuracy is approximately at 55%). Moreover, the GCC-PHAT-CNN shows a slightly better performance in Babble and Destroyerops noise, and its result in WGN noise increases rapidly with increase in the diameter, which is over 90% when d is up to 0.1 m. This shows that the array size has a significant impact on the performance by the GCC-PHAT features, with a larger diameter usually giving better estimation results.

4) Effect of Microphone Mismatches: Fig. 14 shows the localization accuracy of each method when λ is varied from 0 to 1 with a step size of 0.2, i.e., gradually increase in microphone mismatches, where $RT_{60} = 300$ ms and SNR = 10 dB. From Fig. 14, we can see that, the proposed CH-E-MMP-CNN method provides superior performance with the



(a) (b) (c)

Fig. 13. Effect of array size on the performance of each method for localization accuracy with $RT_{60} = 300$ ms and SNR = 10 dB: (a) Babble; (b) Destroyerops; (c) WGN.

increase in microphone mismatches, and the proposed CH-MMP-CNN also achieves stable results. This is probably because the microphone mismatches may affect the value of power spectrum peak, but not the correspondence between the spectrum peak and DOA. In contrast, the proposed ECH-CNN performs less well, due to the gain errors of the microphone. The IRM-BLSTM and PSM-BLSTM approaches are sensitive to the Destroyerops noise, as aforementioned, giving the lowest performance at around 30%. Moreover, the IRM-BLSTM and PSM-BLSTM perform well and offer similar accuracy for Babble noise (around 95%) and WGN (around 85%). The STFT-CNN and GCC-PHAT-CNN algorithms are outperformed slightly by the proposed methods for Babble and Destroyerops noises, but considerably for WGN noise, with the accuracy dropped to nearly 30%. The gain errors from the microphone signals may be the main reason for the inaccurate DOA estimation results by the STFT-CNN method, while the phase errors in the cross-spectrum may be the main reason for inaccurate localization results by the GCC-PHAT-CNN method.

5) Effect of Elevation: Fig. 15 shows the localization accuracy of our proposed methods when the elevation is 30°, 60° and 90°, respectively, with the reverberation time RT_{60} varying from 200 to 800 ms and the level of noise in terms of SNR is 10 dB. Fig. 16 shows the localization accuracy with the SNR varying from -5 to 20 dB and the reverberation

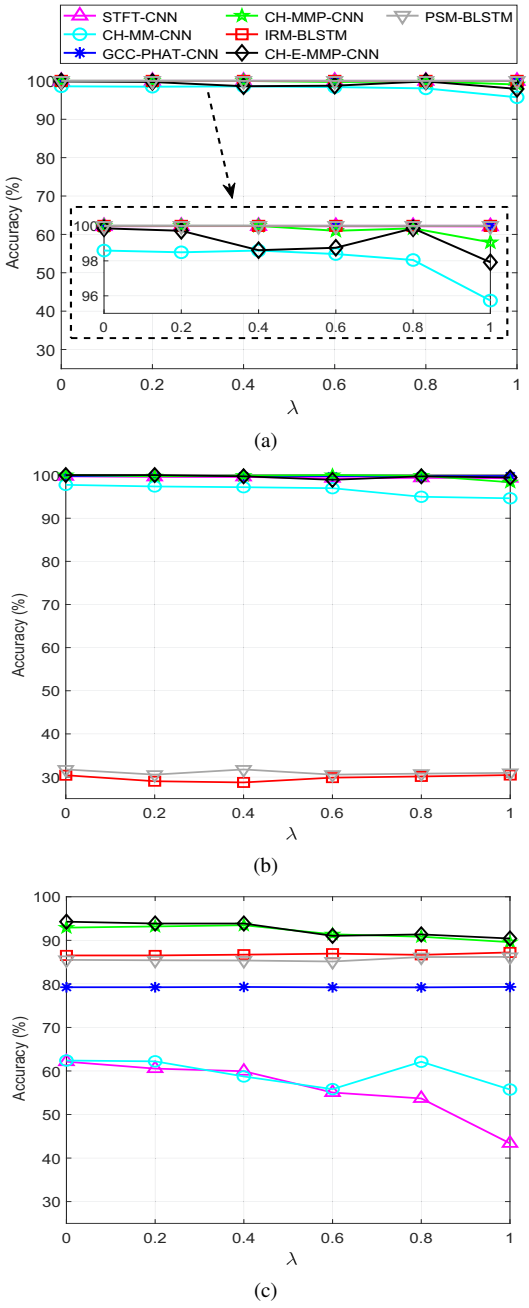


Fig. 14. Effect of microphone mismatches on the performance of each method for localization accuracy with $RT_{60} = 300$ ms and SNR $= 10$ dB: (a) Babble; (b) Destroyerops; (c) WGN.

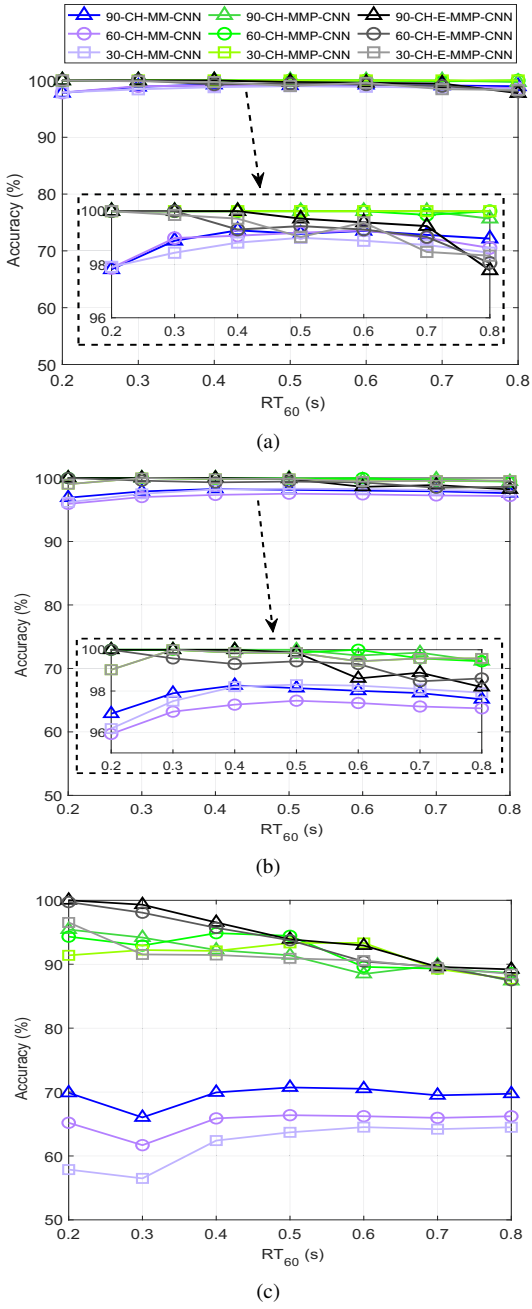


Fig. 15. Effect of elevation on the performance of proposed methods for localization accuracy in the fixed-room with various RT_{60} and SNR $= 10$ dB: (a) Babble; (b) Destroyerops; (c) WGN.

time at $RT_{60} = 300$ ms. Herein, we only consider a fixed room of size 9.7 m $\times$ 7.05 m $\times$ 3 m, to evaluate the effect of different elevations. From Fig. 15 and Fig. 16, we can see that the three proposed methods show similar performance trends in reverberant and noisy conditions. Specifically, the CH-E-MMP-CNN approach still outperforms other two proposed algorithms. This result indicates the effectiveness of the circular harmonic enhanced function and the reliability of this proposed method. Likewise, the localization accuracy of the proposed CH-MMP-CNN method is similar to that of the CH-E-MMP-CNN method, under the various levels of SNR and RT_{60}. This demonstrates the effectiveness of modes magnitude and phase based features. However, the proposed ECH-CNN method gives the lowest localization performance among three proposed algorithms, especially when the elevation is 30°, the accuracy is at 10% for SNR $= -5$ dB in WGN noise.

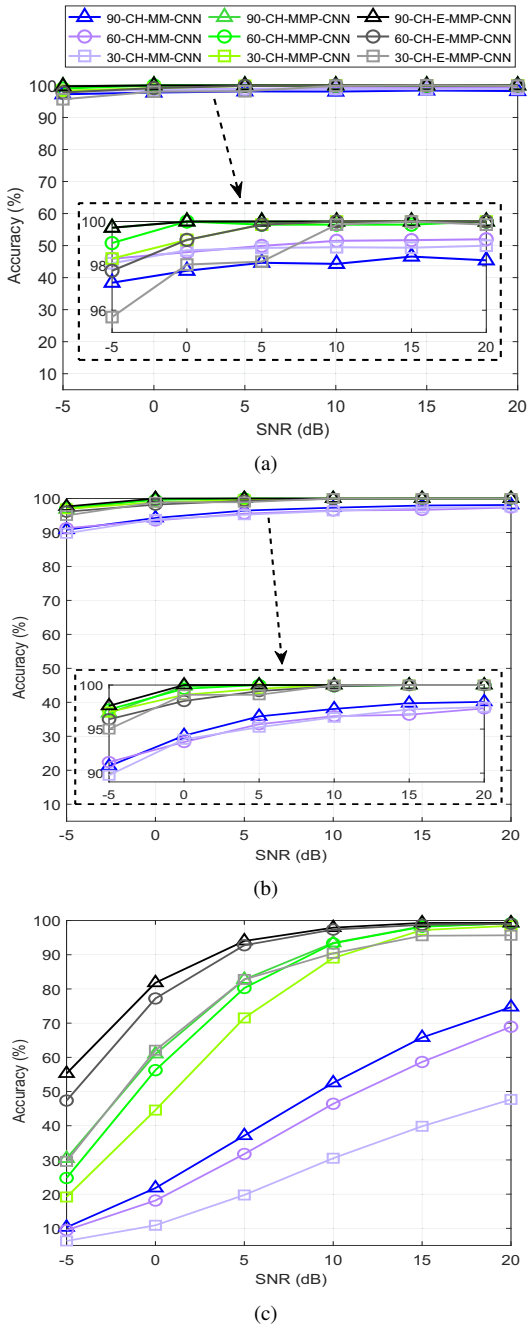


Fig. 16. Effect of elevation on the performance of the proposed methods in localization accuracy in the fixed-room with various SNR and $RT_{60} = 300$ ms: (a) Babble; (b) Destroyerops; (c) WGN.

6) Effect of Temperature: We would like to remark that the temperature only affects the speed of sound, thereby the choice of the order N. When the indoor temperature is 15°C as we considered in our experiments, the speed of sound is 340 m/s. When the indoor temperature is 10°C, the speed is 337 m/s, and when the indoor temperature is 20°C, the speed is 343 m/s. Due to the small change in sound speed, the impact on the choice of the order N is negligible. For example, if the radius $r = 0.02$ m, $f = 3500$ Hz, $M = 4$, the order would be $N = 1$ for all the three sound speeds. This means that the temperature changes would not be an issue of concern for our methods.

F. Source Localization Results in Real-World Experiments

Table III shows the localization accuracy of various methods when the spatial resolution is 20°. As can be observed, results on real data are generally consistent with the aforementioned results on synthetic data. For most of the DOAs, the proposed CH-E-MMP-CNN algorithm provides the highest average localization accuracy, up to 97.50%, which indicates the effectiveness of using the circular harmonic enhanced function in a practical environment. The proposed CH-MMP-CNN method shows slight degradation in accuracy, however, it performs better than CH-E-MMP-CNN for some DOAs. Furthermore, similar to the findings in the synthetic simulation, the IRM-BLSTM and PSM-BLSTM approaches perform well, giving average accuracy over 83%. The proposed ECH-CNN, GCC-PHAT-CNN and STFT-CNN methods give the average localization accuracy below 70%, which are inferior to those of the three methods mentioned earlier.

To further evaluate our methods, we used the Task 1 of the LOCATA challenge [51], [52], which aims to localize a single and static loudspeaker via using a spherical array from the Eigenmike manufactured by mh acoustics [53]. Specifically, we choose the microphone number 6, 12, 22, and 28 of the Eigenmike to emulate the circular microphone, and use the same parameter setup aforementioned. The localization accuracy achieved on this dataset is STFT-CNN: 35.23%, GCC-PHAT-CNN: 54.55%, IRM-BLSTM: 72.73%, PSM-BLSTM: 77.27%, ECH-CNN: 50.44%; CH-MMP-CNN: 81.82%; CH-E-MMP-CNN: 86.36%, respectively. These results further verify the effectiveness of our proposed CH-MMP-CNN and CH-E-MMP-CNN methods.

VII. Conclusion

We have presented a new acoustic source localization approach using deep learning architecture in the circular harmonic domain. The novel contributions of our work are on the new features designed in circular harmonic domain. This enables a CNN framework to be applied to the microphone signals to learn a mapping from the acoustic features to the DOAs of the sources. Our approach achieves competitive performance (e.g. accuracy and stability) in DOA estimation under a variety of array size, noise and reverberation conditions (including conditions unseen in the training stage). Numerical results on

TABLE III
THE LOCALIZATION ACCURACY OF THE EVALUATED METHODS IN THE REAL EXPERIMENTS, WHERE THE SPATIAL RESOLUTION IS 20°. THE RESULTS WITH THE HIGHEST ACCURACY ARE HIGHLIGHTED IN BOLD.

DOA (deg)	STFT-CNN	GCC-PHAT-CNN	IRM-BLSTM	PSM-BLSTM	ECH-CNN	CH-MMP-CNN	CH-E-MMP-CNN
−170°	45%	55%	80%	**95%**	60%	90%	**95%**
−150°	50%	70%	70%	85%	65%	**100%**	**100%**
−130°	75%	90%	65%	85%	55%	95%	**100%**
−110°	65%	85%	80%	75%	50%	90%	**95%**
−90°	65%	75%	**100%**	95%	65%	**100%**	**100%**
−70°	40%	60%	80%	70%	60%	90%	**100%**
−50°	35%	40%	70%	90%	55%	**100%**	**100%**
−30°	30%	40%	80%	**100%**	65%	90%	**100%**
−10°	70%	80%	**100%**	90%	70%	**100%**	**100%**
10°	60%	70%	85%	90%	55%	**95%**	**95%**
30°	80%	100%	100%	95%	70%	**100%**	**100%**
50°	50%	80%	70%	**100%**	80%	90%	**100%**
70°	70%	90%	**100%**	95%	65%	95%	**100%**
90°	70%	90%	95%	90%	50%	**100%**	85%
110°	45%	60%	90%	85%	55%	90%	**100%**
130°	50%	60%	95%	**100%**	55%	**100%**	**100%**
150°	60%	60%	80%	70%	65%	**100%**	**100%**
170°	30%	40%	55%	60%	50%	**85%**	**85%**
Average	55.00%	69.17%	83.33%	85.56%	60.56%	94.44%	**97.50%**

simulated and real room environments demonstrated the state-of-the-art performance of our proposed approaches as compared with several recent baseline methods. In the future, we will extend the proposed methods to multi-source scenarios.

VIII. ACKNOWLEDGEMENT

The authors wish to thank the associate editor and the anonymous reviewers for their helpful suggestions. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

REFERENCES

[1] H. Purwins, B. Li, T. Virtanen, J. Schlter, S. Chang and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206-219, May, 2019.

[2] E. Bezzam, R. Scheibler, J. Azcarreta, H. Pan, M. Simeoni, R. Beuchat, P. Hurley, B. Bruneau, C. Ferry and S. Kashani, "Hardware and software for reproducible research in audio array signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 6591-6592.

[3] M. Yasuda, Y. Koizumi, L. Mazzon, S. Saito, and H. Uematsu, "DOA estimation by DNN-based denoising and dereverberation from sound intensity vector," *arXiv e-prints: 1910.04415*, Oct. 2019.

[4] W. Manamperi, T. D. Abhayapala, J. Zhang and P. N. Samarasinghe, "Drone audition: sound source localization using on-board microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 508-519, Jan. 2022.

[5] N. Ma, T. May and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2444-2453, Dec. 2017.

[6] B. Laufer-Goldshtein, R. Talmon, I. Cohen and S. Gannot, "Multi-view source localization based on power ratios," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 71-75.

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

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

[9] X. Yue, G. Qu, B. Liu and A. Liu, "Detection sound source direction in 3D space using convolutional neural networks," in *Proc. First Int. Conf. AI Ind. (AI4I)*, Laguna Hills, CA, Sep. 2018, pp. 81-84.

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

[11] H. Sundar, T. V. Sreenivas and C. S. Seelamantula, "TDOA-based multiple acoustic source localization without association ambiguity," *IEEE/ACM Trans. Audio, Speech, Lang. Process.,* vol. 26, no. 11, pp. 1976-1990, Nov. 2018.

[12] Z.-Q. Wang, X. Zhang, and D. L. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks, in *Proc. of Interspeech*, Hyderabad, India, Sept. 2018, pp. 322-326.

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

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

[15] S. Adavanne, A. Politis and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Dec. 2018, pp. 1462-1466.

[16] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," *Ph.D. dissertaion*, Brown Univ., Providence, RI, USA, 2000.

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

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

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

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

[21] H. Teutsch, *Modal Array signal processing: principles and applications of acoustic wavefield decomposition*, Berlin/Heidelberg, Germany: Springer-Verlag, 2007.

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

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

[24] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path

dominance test," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494-1505, Oct. 2014.
[25] K. SongGong and H. Chen, "Robust indoor speaker localization in the circular harmonic domain," *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3413-3422, Apr. 2021.
[26] K. SongGong, H. Chen and W. Wang, "Indoor multi-speaker localization based on bayesian nonparametrics in the circular harmonic domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1864-1880, May, 2021.
[27] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 2814-2818.
[28] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, May, 2016, pp. 405-409.
[29] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Apps. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Dec., 2017, pp. 136-140.
[30] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8-21, Mar., 2019.
[31] L. Perotin, R. Serizel, E. Vincent and A. Gurin, "CRNN-based joint azimuth and elevation localization with the ambisonics intensity vector," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Tokyo, Japan, Nov., 2018, pp. 241-245.
[32] L. Perotin, R. Serizel, E. Vincent and A. Gurin, "CRNN-based multiple DOA estimation using acoustic intensity features for ambisonics recordings," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 22-23, Mar., 2019.
[33] Z. Wang, X. Zhang and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 178-188, Jan. 2019.
[34] Z. Liu, C. Zhang and P. S. Yu, "Direction of arrival estimation based on deep neural networks with robustness to array imperfections," *IEEE Trans. Antennas Propag.*, vol. 66, no. 12, pp. 7315-7327, Dec. 2018.
[35] L. Perotin, A. Défossez, E. Vincent, R. Serizel and A. Guérin, "Regression versus classification for neural network based audio source localization," in *Proc. IEEE Workshop Apps. Signal Process. Audio Acoust. (WASPPA)*, New Paltz, NY, USA, Oct. 2019, pp. 343-347.
[36] H. Teutsch and W. Kellermann, "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2724-2736, Aug. 2006.
[37] A. M. Torres, J. Mateo, and M. Cobos, "Room acoustics analysis using circular arrays: A comparison between plane-wave decomposition and modal beamforming approaches," *Circuits. Syst. Signal Process.*, vol. 35, no. 5, pp. 1625-1642, May 2016.
[38] V. Varanasi, A. Agarwal and R. M. Hegde, "Near-field acoustic source localization using spherical harmonic features," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2054-2066, Dec. 2019.
[39] E. Tiana-Roig, F. Jacobsen, and E. F. Grande, "Beamforming with a circular microphone array for localization of environmental noise sources," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3535-3542, Dec. 2010.
[40] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger and S. Gannot, "Multi-microphone speaker separation based on deep DOA estimation," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, A Coruna, Spain, Sept. 2019, pp. 1-5.
[41] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, New Orleans, LA, USA, Jun., 2017, pp. 6125-6129.
[42] J. Lee, I. Lee, and J. Kang, "Self-attention graph pooling," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, Jun. 2019, pp. 3734-3743.
[43] J. Sun, Convolutional neural networks, 2017. [Online]. Available: https://sites.cc.gatech.edu/ san37/post/dlhc-cnn/.
[44] P. A. Grumiaux, S. Kitiá, L. Girin, and A. Guárin, "A survey of sound source localization with deep learning methods", arXiv:2109.03465v2, Sep. 2021.
[45] I. Goodfellow, Y. Bengio and A. Courville, *Deep learning*, Massachusetts: MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org
[46] E. A. P. Habets, "RIR Generator," 2016. [Online]. Available: https://github.com/ehabets/RIR-Generator
[47] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943-950, Apr. 1979.
[48] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247-251, Jul. 1993.
[49] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," National Institute of Standards and Technology, Gaithersburg, MD, USA, NIST Interagency/Internal Rep. 4930, Feb. 1993.
[50] Q. Li, X. Zhang and H. Li, "Online direction of arrival estimation based on deep learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 2616-2620.
[51] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking, in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop (SAM)*, Sheffield, U.K., Jul. 2018, pp. 410-414.
[52] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, IEEE-AASP challenge on acoustic source localization and tracking-documentation of final release (Version 1.0), Jan. 2020. [Online]. Available: www.locata-challenge.org.
[53] mh acoustics, EM32 Eigenmike microphone array release notes (v17.0), Oct. 2013. [Online]. Available: www.mhacoustics.com/sites/default/files/ReleaseNotes.pdf.

Kunkun SongGong received the B.S. degree from Hefei Normal University, China, in 2012, the M.S. degree from the Nanjing University of Information Science and Technology, China, in 2016, and the Ph.D. degree from Nanjing University of Aeronautics and Astronautics, China, in 2022. He was a visiting Ph.D. student with the Centre for Vision Speech and Signal Processing, University of Surrey, UK, in 2021.

Wenwu Wang is a Professor in Signal Processing and Machine Learning, and a Co-Director of the Machine Audition Lab within the Centre for Vision Speech and Signal Processing, University of Surrey, UK. He is also an AI Fellow at the Surrey Institute for People Centred Artificial Intelligence. He is a Senior Area Editor for IEEE Transactions on Signal Processing, an Associate Editor for IEEE/ACM Transactions on Audio Speech and Language Processing, an elected Vice Chair of IEEE Machine Learning for Signal Processing Technical Committee, and an elected Member of the IEEE Signal Processing Theory and Methods Technical Committee.

Huawei Chen received the Ph.D. degree in underwater acoustic engineering from Northwestern Polytechnical University, Xian, China, in April 2004. From August 2005 to August 2009, he was with the Centre for Signal Processing, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, as a Research Fellow. Since September 2009, he has been a Full Professor with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests include microphone array signal processing, acoustical and speech signal processing, statistical and adaptive signal processing. He is an Associate Editor of Circuits, Systems, and Signal Processing.