MULTI-SPEAKER LOCALIZATION IN THE CIRCULAR HARMONIC DOMAIN ON SMALL APERTURE MICROPHONE ARRAYS USING DEEP CONVOLUTIONAL NETWORKS

Kunkun SongGong[1]*, Pufen Zhang*[1*]*, Xiongwei Zhang*[1]*, Meng Sun*[1]*, and Wenwu Wang*[2]

[1]Army Engineering University, China
[2]Centre for Vision, Speech and Signal Processing, University of Surrey, U.K.
Email: sgkk@nuaa.edu.cn, pufenzh@163.com, xwzhang9898@163.com, sunmeng@aeu.edu.cn, w.wang@surrey.ac.uk

ABSTRACT

Acoustic signal processing in the circular harmonic domain (CHD) is an appealing method for speaker localization, since it inherently supports wideband acoustic sources and provides frequency invariant beampatterns. However, the performance of existing circular harmonic direction-of-arrival (DOA) estimation approaches can be degraded by a variety of factors, including background noise and reverberation in the acoustic environments, small aperture size of the circular array and the presence of multiple active sources. This paper addresses these issues by proposing a novel multi-speaker CHD localization method with small-sized microphone arrays using deep convolutional neural networks (CNN). The core idea is to construct circular harmonic features through joining the selected time-frequency (TF) bins of higher power and the operation of a randomization process by mimicking the sparsity property of speech signals. After that, we implement multi-speaker estimation as a multi-label classification task, and propose to use CNN with binary cross-entropy as the loss function. Experimental results show that our method performs significantly better than the baseline methods, on both simulated and real data, in terms of the accuracy of DOA estimation.

Index Terms— Multi-speaker localization, circular harmonic, deep convolutional networks, multi-label classification

1. INTRODUCTION

Acoustic source localization (ASL) is the problem of estimating the position of single or multiple sound sources, relative to the position of recording microphone array. In most cases, ASL is simplified to estimate the DOAs of the sound sources, i.e., the azimuth or/and elevation angles of these sources [1, 2]. ASL is crucial for a variety of practical applications, such as smart home, robotics, among many others [3–5]. In the literature, DOA estimation methods can be broadly classified into five categories: 1) the time difference of arrival (TDOA) of sound sources, e.g. by exploiting the Generalized Cross Correlation PHAse Transform (GCC-PHAT) [6], 2) the subspace-based approaches, including the popular methods e.g. multiple signal classification (MUSIC) [7] and estimation of signal parameters via rotational invariance techniques (ESPRIT) [8], 3) the beamforming based approaches, e.g. steered response power with phase transform (SRP-PHAT) [9], 4) the intensity-based methods, e.g. via determining the magnitude and direction of the transport of acoustic energy, related to the DOA of a sound wave [10], and 5) the emerging ASL approaches based on data-driven deep neural networks (DNNs) technique, which utilize the variety of input features, e.g. inter-channel features [11], cross correlation-based features [12], spectrogram-based features [13], ambisonic signal representation based features [14], intensity-based features [15] and waveforms [16].

In the past several years, the circular harmonic DOA estimation technique has received increasing attention, since it can provide a frequency-invariant eigen-beam pattern that is useful for localizing wideband sources without the narrowband assumption used by traditional signal models [5]. However, the performance of this method degrades in high level of background noise and room reverberation, for multiple speakers, and in the presence of small-sized arrays.

On the basis of our prior work for studying the single source localization based on CNN [17], in this paper, we proposed a novel CHD multi-speaker localization method on small-sized microphone arrays using deep convolutional networks. Our objective is to localize multiple sound sources in the CHD by learning the mapping from the acquired sensor signals to DOA of sources using a large set of training data with multiple labels, using CNN with a binary cross-entropy (BCE) loss. First, we construct the circular harmonic features through joining the selected time-frequency (TF) bins of higher power and the operation of a randomization process by exploiting the sparsity property of speech signals, which can reduce the adverse impact of noise, reverberation and multi-source. Then, we leverage the advantage of DNNs in generalizing to diverse scenarios and array geometries, for example the small aperture arrays, to further improve the localization performance. Through experimental evaluations, we show the superior performance of the proposed method.

2. SIGNAL MODEL

A uniform circular array (UCA) consisting of M omnidirectional sensors is adopted for sound source capture as shown in Fig. 1. The radius of the array is r, and the azimuth angle of each sensor is ϑ_m, namely,

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

Suppose that D far-field speech sources in a reverberant enclosure impinge on the array. Herein, the DOAs are defined with respect to the positive x-axis, which implies $\theta_d \in [-\pi, \pi), d = 1, \ldots, D$. In the short-time Fourier

This work was supported by the National Natural Science Foundation of China (62071484, 62371469).

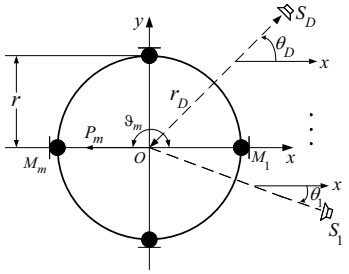


Fig. 1. Configuration of the uniform circular sensor array.

transform (STFT) domain, the source signals received at the mth sensor can be modeled as

$$P_m(k,t) = \sum_{d=1}^{D} H_{dm}(k,t)S_d(k,t) + V_m(k,t), \quad (2)$$

where $k = 2\pi f/c$ is the wavenumber, t is the time frame index, f is the frequency and c is the speed of sound. $S_d(k,t)$ is the signal induced by one of the D speech sources at r_d distance from the centre O of the microphone array, $H_{dm}(k,t)$ is the room impulse responses (RIRs) from the dth source to the mth sensor, and $V_m(k,t)$ is the additive background noise.

Since speech signals are considered sparse in the TF domain, at each TF bin, it could be assumed that only one source is dominant [18]. Thus, (2) can be further simplified as

$$P_m(k,t) \approx H_{dm}(k,t)S_d(k,t) + V_m(k,t). \quad (3)$$

3. PROPOSED METHOD

3.1. Circular Harmonic Beamforming (CHB)

The aim of CHB is to combine different harmonic components to form a beam with appropriate spatial selectivity properties. In real-life applications, the discretization of the continuous aperture by means of a uniform circular array with M omnidirectional sensors leads to the nth-order circular harmonic beam response [3, 19], as follows

$$\begin{aligned} B(k,t) &= \sum_{n=-N}^{N} \tilde{C}_n(k,t) \cdot G_n(k) \cdot H_n(\theta) \\ &= \sum_{n=-N}^{N} \frac{1}{M} \sum_{m=1}^{M} P_m(k,t) e^{-jn\vartheta_m} \cdot \frac{1}{j^n J_n(kr)} \cdot e^{jn\theta}, \end{aligned} \quad (4)$$

where $\theta \in [-\pi, \pi)$, $j = \sqrt{-1}$, and n is the order of harmonic. $J_n(\cdot)$ is the nth-order Bessel function of the first kind, $\tilde{C}_n(k,t)$ represents the circular harmonics, G_n is an equalization factor, and H_n is a frequency-independent phase factor [20]. Thus, Equation (4) forms the basis of the proposed circular harmonic features discussed in the ensuing sections. Note that in practice, the number of harmonics must be truncated to a maximum order N, which is related to the number of sensors [20], 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.

3.2. The Circular Harmonic Feature based on Selected and Randomized Processing

Since the baseline TF-CHB method is sensitive to background noise, reverberation and the characteristics of the array, particularly the small aperture array, this may degrade the performance of the DOAs estimation in adverse environments. To address this issue, we propose a novel circular harmonic feature based on selected and randomized processing.

Selected Processing: In our earlier work [5], 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. Herein, 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, to help us find the useful TF bins with high power in the CHD.

The circular harmonic feature based on selected processing, which contains I components, can be represented as

$$\mathbf{F}(k,t,\theta) = \left[B^E(k,t,\theta_1), \cdots, B^E(k,t,\theta_i), \cdots, B^E(k,t,\theta_I) \right]^T \quad (5)$$

where

$$B^E(k,t,\theta_i) = \sum_{n=-N}^{N} E_f(k,t) \cdot \tilde{C}_n(k,t) \cdot G_n(k) \cdot H_n(\theta_i), \quad (6)$$

and the number of classes I depends on the resolution of the whole range of DOAs, and θ_i is the DOA corresponding to the ith class. The selected function $E_f(k,t)$ can be represented as

$$E_f(k,t) = \begin{cases} 1, & \text{if } E_0(k_q,t_q) \geq E_0(k_u,t_u) \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

where $E_0(k,t) = \left| \tilde{C}_0(k,t) \cdot G_0(k) \cdot H_0(\theta_i) \right|^2$ is the power of the 0-th mode strength. $E_0(k_u,t_u)$ is the power at the uth TF bin, $u = \alpha Q$, with $\alpha \in (0,1]$ being a pre-defined threshold, e.g. $\alpha = 0.9$. Q stands for the total number of TF bins. $E_0(k_q,t_q)$ is the power at the qth TF bin.

Randomized Processing: With the sparsity property of speech as aforementioned, we introduce a randomization process to exploit the non-overlapping and sparsity property of speech signals [21]. More specifically, we randomly assign the activity of each sub-band source across different frequency bands to ensure that each TF bin corresponds to the activity of a separate source, with different sources being active in different TF bins. Consequently, the resulting training signal more realistically reflects the sound distribution and characteristics of a multi-source environment, allowing the neural network to effectively learn relevant TF-CHB-Selected-Randomized feature (for simplicity, TF-CHB-S-R) for localization.

As an example, we consider two speakers. The randomization process is conducted as follows. First, for a given small-sized array setup, the selected TF-CHB representation of multi-microphone signals, corresponding to two different angles of speakers, are concatenated along the time frame. Then, for each sub-band, the TF bins across

Table 1. Changeable-room configurations in training process and test process.

Array: M=4, r=0.02 m; DOAs: $-120^\circ, 80^\circ$; Sound speed: 340 m/s; TIMIT speech signals: 0.5 s length; Sampling rate: 16 kHz; Noise types [23]: training process – Babble, Factory2 and Volvo, test process – Destroyerops and Hfchannel.

Reverberant Conditions				
Type	Length(m) × Width(m) × Height(m)	Distance(m)	RT_{60}(ms)	SNR(dB)
Train	Length: picked randomly between 7 and 12; Width: picked randomly between 5 and 9; Height: picked randomly 3 and 4;	picked randomly 1.5 and 2	picked randomly between 200 and 800	20
Test	$9.7 \times 7.05 \times 3.0$	1.5	200 to 800 with an increment of 100	20
Noisy Conditions				
Type	Length(m) × Width(m) × Height(m)	Distance(m)	SNR(dB)	RT_{60}(ms)
Train	Length: picked randomly between 7 and 12; Width: picked randomly between 5 and 9; Height: picked randomly 3 and 4;	picked randomly 1.5 and 2	picked randomly between 0 and 30	200
Test	$9.7 \times 7.05 \times 3.0$	1.5	0 to 30 with an increment of 5	200

all the microphones are randomized, resulting in a single multi-channel signal [21]. This process ensures that the number of TF bins that contain activities related to the two desired DOAs is approximately the same for each time frame. Note that, the dimension of the input to the network is $I \times N_{fb}$ (number of frequency bins) for the proposed TF-CHB-S-R. The selected and randomized processing is summarized in Algorithm 1.

Algorithm 1 The Selected and Randomized Processing

Input: $P_m(k,t)$, $E_f(k,t)$, u.

1: **for** each $g \in I$ **do**
2: using (6) to obtain $B_g^E(k_u, t_u, \theta_g)$ for DOA θ_g
3: **for** each $h \in I$ and $h \neq g$ **do**
4: using (6) to obtain $B_h^E(k_u, t_u, \theta_h)$ for DOA θ_h
5: B_{gh}^E : concatenate B_g^E with B_h^E along time axis and randomize TF bins
6: calculate the TF-CHB-S-R feature, i.e. the $I \times N_{fb}$ matrix for each time frame u from B_{gh}^E
7: **end for**
8: **end for**

3.3. CNN Network Architecture

Table 2. The architecture of the CNN for multi-speaker localization.

L1 (Layer 1): Input	$I \times N_{fb}$	L6: Conv 3	64 filters @2× 1	L11: ReLU 4	-
L2: Convolutional Layer 1 (Conv 1)	64 filters @2× 1	L7: ReLU 3	-	L12: FC 2	512
L3: ReLU 1	-	L8: Dropout 1	0.5	L13: Dropout 3	0.5
L4: Conv 2	64 filters @2× 1	L9: Fully Connected 1 (FC 1)	512	L14: ReLU 5	-
L5: ReLU 2	-	L10: Dropout 2	0.5	L15: Output	I× 1

In this study, we choose the CNNs as the network for the multi-DOAs classification purpose. The architecture of the detailed CNN is illustrated in Table 2. The stochastic gradient descent with momentum (SGDM) is chosen as the optimizer. Dropout with a rate 0.5 is used to mitigate overfitting. The size of mini-batches is set as 4, 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. To address the multi-label classification problem, we use the binary relevance method. The output layer of the CNN consists of I sigmoid activations, each corresponding to a DOA class, using BCE as the loss function. Thus, the DOA estimation approaches based are TF-CHB-SR-CNN, TF-CHB-S-CNN and TF-CHB-R-CNN. The latter two are the methods using the proposed features TF-CHB-Selected and TF-CHB-Randomized, respectively.

4. EXPERIMENTAL EVALUATIONS

4.1. Datasets and Set up

The simulated data used for training were generated in a changeable-room and the detailed configurations are shown in Table 1. Herein, the presented objective evaluations were for the two speakers scenario. Please note that, by following the same procedure as aforementioned, the method can be extended for estimating more than two speakers scenarios. In the simulations, 100 utterances were randomly selected for training, and another 20 utterances for testing. Thus, for the reverberant condition, we used $100 \times (18 \times 17) \times 7 \times (6 \times 5 \times 2) \times 2 \times 3 = 77112000$ training signals, where 100 represents the number of utterances, 18×17 represents the number of candidate DOAs, 7 represents the number of reverberant conditions with reverberation time varied from 200 to 700ms, $6 \times 5 \times 2$ represents the number of the room size, 2 represents the number of sources to array distances, and 3 represents the number of noise types. For the noise condition, the number of training signals is also 77112000, where we consider 7 noise levels with signal to noise ratio varied from 0 to 30 dB. The length of speech signals is 0.5 s. We considered that the spatial resolution for source location was set to be 20° in the simulations. Thus, the total number of candidate source locations was $I = 18$, which were distributed uniformly from -180° to 160°. To generate the RIRs [22] 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. The performance of the proposed TF-CHB-S-CNN, TF-CHB-R-CNN and TF-CHB-S-R-CNN are evaluated and compared with the baselines including the STFT-CNN [21] and the TF-CHB-Histogram [19] methods in both simulated and real environments. 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 $N_{fb} = 511$.

The reason for using 511 bins is that the DC component and Nyquist frequency are neglected as they do not provide any useful localization information.

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 RT_{60} of approximately 350 ms. The two speakers were located at two DOAs of -120° and 80°, respectively, and other conditions were similar to those in the above simulations. 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. A photograph of the real experiment and microphone array is shown in Fig. 2.

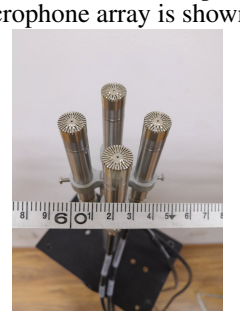


Fig. 2. The uniform circular sensor array used in the experiments, with a array radius of 0.02 m.

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

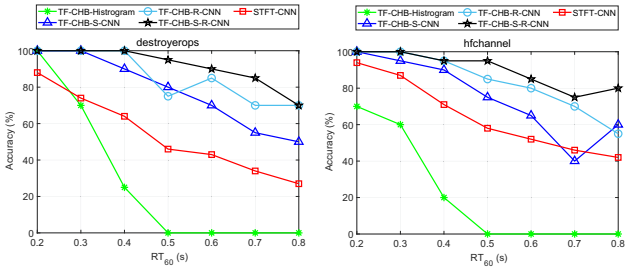
$$Acc = \frac{N_{cr}}{N_e} \times 100\%, \quad (8)$$

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 $\pm 20^\circ$.

4.2. Results in Simulated Experiments

Effect of Room Reverberation: Fig. 3 shows the localization accuracy of each method under the conditions of the changeable-room with the RT_{60} varying. In general, the performance of all tested algorithms degrades with the increase in the level of reverberation. The proposed TF-CHB-S-R-CNN outperforms all the baseline methods. With the selected and randommized processing, more reliable TF bins, which
are dominated by the multi-sources, are learned by the CNN, leading to improved localization performance. In contrast, the TF-CHB-Histogram method performs poorly, especially under higher reverberation conditions, and have failed to localize the sources when RT_{60} is over 500 ms. The main reason is the small sized-array and the presence of strong reverberations, which significantly degrade localization results. We can also see that our method outperforms the TF-CHB-S-CNN, the TF-CHB-R-CNN and the STFT-CNN methods with various reverberant environments.

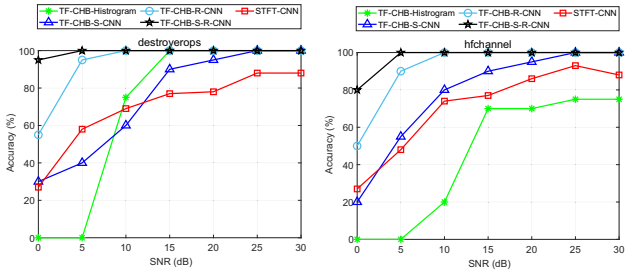
Effect of Noise Level: Fig. 4 shows the localization accuracy of each method under the changeable-room conditions with SNR varying. The proposed TF-CHB-S-R-CNN method



(a) Destroyerops (b) Hfchannel

Fig. 3. Effect of room reverberation on the performance of each method for localization accuracy in the changeabled-room with SNR $= 20$ dB.

offers better localization performance under both noise types as compared with the baseline methods. This is because the proposed approach can select the TF bins which are less affected by noise, and improve the estimation accuracy of muliti-speaker with the help of randomized processing. However, regarding the robustness, the TF-CHB-Histogram method shows higher variances over noisy conditions. In between are the remaining methods, namely the TF-CHB-S-CNN, the TF-CHB-R-CNN and the STFT-CNN.



(a) Destroyerops (b) Hfchannel

Fig. 4. Effect of noise level on the performance of each method for localization accuracy in the changeable-room with $RT_{60} = 200$ ms.

4.3. Results in Real-World Experiments

The localization accuracy achieved on the real dataset is STFT-CNN: 35%, TF-CHB-Histogram: 35%, TF-CHB-S-CNN: 50%, TF-CHN-R-CNN: 70% and TF-CHB-S-R-CNN: 85%, respectively. As can be seen, the localization results behave in a similar manner to those found in the simulation results above. The results show good performance of our proposed TF-CHB-S-R-CNN method, which indicates the effectiveness of using the circular harmonic feature based on the selected and randomized processing in a practical environment.

5. CONCLUSION

We have presented a TF-CHB-S-R-CNN model to estimate multi-source with a small-sized array in adverse environments. Our contribution is on the way of designing and improving the circular harmonic feature via selected and randomized processing, which enables a CNN to be utilized for multi-DOA estimation in noisy and reverberant conditions. The experimental results demonstrated that the proposed method offers better performance than the compared methods.

6. 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] PA. Grumiaux, S. Kitic, L. Girin and A. Guerin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Am.*, vol. 152, no. 1, pp. 107-151, Jul. 2022.

[3] 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.

[4] F. Grondin, M. A. Maheux, J. S. Lauzon, J. Vincent and F. Michaud, "Fast cross-correlation for TDOA estimation on small aperture microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1-5.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] M. Cobos, M. Pezzoli, F. Antonacci and A. Sarti, "Acoustic source localization in the spherical harmonics domain exploiting low-rank approximations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1-5.

[11] S. Gannot, E. Vincent, S. Markovich-Golan and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692-730, Apr. 2017

[12] L. Comanducci, F. Borra, P. Bestagini, F. Antonacci, S. Tubaro and A. Sarti, "Source localization using distributed microphones in reverberant environments based on deep learning and ray space rransform," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2238-2251, Jul. 2020.

[13] G. Bologni, R. Heusdens, and J. Martinez, "Acoustic reflectors localization from stereo recordings using neural networks, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 1-5.

[14] S. Adavanne, A. Politis, J. Nikunen and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent meural metworks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34-38, Mar., 2019.

[15] Liu, N., Chen, H., Songgong, K., and Li, Y, "Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays, *J. Acoust. Soc. Am.*, vol. 149, no. 2, pp. 1069-1084, Feb. 2021.

[16] T. N. Sainath et al., "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 965-979, May, 2017.

[17] K. SongGong, W. Wang and H. Chen, "Acoustic source localization in the circular harmonic domain using deep learning architecture," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2475-2491, Jul. 2022.

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

[19] 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.

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

[21] 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.

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

[23] 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.