DOI: 10.61782/fa.2023.1180

LEARNING-BASED MASKING FOR RELIABLE SOURCE LOCALIZATION INTERFERED BY UNDESIRED DIRECTIONAL NOISE

Priyadarshini Dwivedi[1*] **Gyanajyoti Routray** [1] **Siddesh B. Hazare** [1] **Rajesh M. Hegde** [2]

[1] Department of Electrical Engineering, Indian Institute of Technology Kanpur, India.
[2] Department of Electrical Engineering, Indian Institute of Technology Dharwad, India.

ABSTRACT

It is incredibly challenging to simultaneously locate an acoustic source in a noisy, reverberant environment and mitigates directional interference. The proposed study uses a spherical harmonic decomposition method to determine the spherical harmonics phase magnitude (SH-PM) components corresponding to the received spherical microphone array (SMA) signals. Before SH-PM components are used as input features to the CNN model, binary masking removes directional interference and emphasizes the desired audio source. In this work, the binary mask is estimated using the learning technique such that it is possible to reliably discriminate between acceptable and undesired sources using real-time mask estimation. The proposed strategy creates a learning-based mask to enable real-time and reliable filtering of the undesirable source. Because of this, the entire strategy is extremely flexible and adaptable. By creating datasets, extensive simulations evaluate the effectiveness of the offered strategy. Additionally, the approach is experimentally validated by conducting tests in a live lab setting. The significance of the suggested strategy promotes the use of the technique in real-world situations.

Keywords: *Source localization, DOA, convolutional neural network, spherical harmonics, spherical microphone array.*

1. INTRODUCTION

Direction of arrival (DOA) estimation is an important problem in signal processing and has a wide range of applications in fields such as radar, sonar, communication, and audio signal processing. The DOA estimation problem involves estimating the angles of arrival of signals received at an array of sensors. This problem has been studied extensively, and various algorithms have been developed in this context [1]. The DOA estimation algorithms can be broadly classified as time-delay-based [2, 3], beamforming-based, and subspace-based approaches. Time difference of arrival (TDOA) estimates the direction of arrival of the signal from the time difference measurements at which the signal arrives at the spatially separated sensors. The accuracy of TDOA-based DOA estimation depends on the accuracy of the time measurements, which can be affected by various factors such as clock synchronization, signal propagation delays, and measurement noise. Beamforming methods use the spatial filter concept and aim to extract the signal of interest from the received signals by spatially weighting them. The angle of arrival of the signal is estimated by finding the direction in which the spatial filter has the highest output power. Subspace-based methods exploit the signal subspace and noise subspace to estimate the direction of arrival. These methods assume that the signal subspace is orthogonal to the noise subspace and use the eigenvalues and eigenvectors of the covariance matrix of the received signals to estimate the signal subspace and noise subspace. The angle of arrival is then estimated by finding the direction in which the signal subspace has the highest power. MUltiple SIgnal Classification (MUSIC) [4] is a popular subspace-based DOA estimation method that

[*] ***Corresponding author***: *priyadw@iitk.ac.in.*

estimates the angle of arrival by finding the peaks in the pseudo-spectrum of the received signals. These methods mostly use the linear array. Therefore, the DOAs are estimated either in the horizontal or in the vertical plane. DOA estimation using the SMA signals in the spherical harmonics (SH) domain has been investigated to localize the desired sources in the three-dimensional geometry [5]. The spherical harmonics (SH) domain also has several advantages, such as it provides a high spatial resolution in the DOA estimation problem. This is because the spherical harmonics can accurately represent spatial patterns, allowing for a more precise estimation of the direction of arrival of signals. The SH signals are robust to noise and can reduce the computational complexity of DOA estimation. Moreover, there have been significant advancements in using learning-based methods for DOA estimation in recent years. These models can learn complex features and relationships from data and provide high accuracy in DOA estimation. In this context, various learning-based methods have been explored utilizing the SH phase and magnitude as the desired features for training the learning models such convolutional neural network (CNN) in [6, 7], support vector machines (SVM) in [8], and convolutional recurrent neural network (CRNN) in [9]. A high-resolution CNN and matching pursuit model is proposed in [10]. Further, SH signals are analysed, and SH intensity coefficients are explored in [11, 12] for far-field DOA and near-field range estimations.

However, these methods did not consider the effect of any undesired source acting as a directional interference. The DOA estimation gets affected by the presence of noise, and reverberation, along with other sources interfering with the desired source signal. Figure 1 provides a typical illustration of such a scenario. Few methods have been explored to address this, such as attention models employed for DOA estimation in [13–15]. Attention focuses on the frequency bands of desired directional signals [13]. In [14], the speaker beam's attention mechanism is a binary mask focusing on the intended sources' dominant frequency ranges. In addition, a deep neural network's attention-based technique for source separation in [15]. But these methods are limited to the linear arrays. Therefore, [16] considers such cases when an undesired source is present as directional interference, noise, and reverberation. This method explores the SH decomposition of the SMA recordings. Also, in this case, DOA is estimated using a CNN framework, and the mask generation uses the conventional approach.

Thus, in this work, a real-time learning-based mask estimation using the DNN framework along with the CNN-based DOA estimation of only the desired source in the presence of reverberation and diffuse noise is proposed. The proposed approach deals with the DOA estimation in azimuth and elevation direction by considering the SH decomposition of signals received at SMA. A neural network is trained to estimate both the directions. The CNN framework is considered, which learns the proposed features. The CNN learns to map the extracted features corresponding to the class of desired DOA angles. The signal received at the microphone contains the desired sources as well as undesired sources. The approach filters out the features of the received signal when the interfering undesired signals are dominant. Moreover, the DNN-based mask implementation is real-time. It provides a robust and fast method for filtering out the undesired source signal and providing DOA cues only for the desired signal. Therefore the filtered desired spherical harmonics phase magnitude (SH-PM) features act as the input data for training our neural network model, which estimates the accurate DOA in the presence of reverberation and noise. This enhances the accuracy by providing attention to the desired sources and thus improves DOA.

The rest of the paper is organized as follows. The system model, basic definitions, and feature extraction are given in section 2. Dataset generation and the learning framework of the proposed model is presented in section 3. The performance of the proposed model is evaluated in section 4 and section 5 concludes the work.

2. SYSTEM MODEL

This section presents a description of the generic sound field data model in the SH domain, followed by an explanation of the filtering criterion utilized to distinguish sources in the context of source localization when an undesired source is present.

2.1 SH Decomposition

The present research considers an acoustic scene comprising Q source vectors denoted by $\mathbf{s} = [s_1, \ldots, s_Q]^T$, located at radial distances $r_{s1}, \ldots, r_{sQ}$ from the center of the sound scene and oriented in the direction (θ_{sq}, ϕ_{sq}), where $q = 1, \ldots, Q$. The variables (θ, ϕ) are used to represent the elevation and azimuth, respectively. The methodology employed to capture the acoustic scene involves the utilization of an SMA with a radius of r. The SMA is comprised of L microphones arranged at $(\theta_l, \phi_l), l = 1, \ldots, L$. Given that the center of the SMA

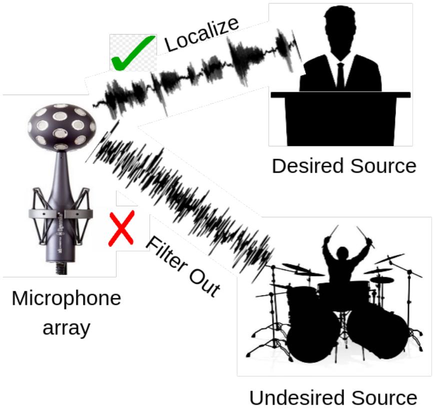


Figure 1: A typical acoustic scenario representing the desired and undesired signal received at the microphone array.

is aligned with the center of the acoustic scene, the sound pressure exerted on the SMA can be expressed as follows

$$\mathcal{P}(k) = \mathbf{Y}(\theta_l, \phi_l)\mathbf{B}(kr)\mathbf{Y}^H(\theta_{sq}, \phi_{sq})\mathbf{s}(k) + \mathbf{z}(k) \quad \forall\, k. \tag{1}$$

The symbol k represents the wave number corresponding to the frequency $\tilde{f}$, where c is the sound velocity. $\mathcal{P}(k) = [p(k, \theta_1, \phi_1), \ldots, p(k, \theta_L, \phi_L)]^T$ and $\mathbf{z}(k) = [z_1(k), \ldots, z_L(k)]^T$ represents the sound pressure vector the noise vector, respectively and $[\cdot]^T$ denotes transpose. $\mathbf{Y}^H(\theta_{sq}, \phi_{sq}) \in \mathbb{C}^{(N+1)^2 \times Q}$ represents the SH matrix for the source positions and given by [17]

$$\mathbf{Y}^H(\theta_{sq}, \phi_{sq}) = [\mathbf{y}_1^H(\theta_{s1}, \phi_{s1}), \ldots, \mathbf{y}_Q^H(\theta_{sQ}, \phi_{sQ})]$$
$$\mathbf{y}(\theta, \phi) = [Y_{00}(\theta, \phi), \ldots, Y_{NN}(\theta, \phi)]^T$$

$[\cdot]^H$ represents the conjugate transpose. The SH basis functions are denoted as $Y_{nm}(\theta, \phi)$ for order $n = 0, \ldots, N$ and degree $m = -n, \ldots, n$. The order of SMA is N and $Y_{nm}(\theta, \phi)$ is given as

$$Y_{nm}(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\phi} \tag{2}$$

$P_n^m(\cdot)$ denotes the associated Legendre polynomial of order n and degree m. Matrix $\mathbf{Y}(\theta_l, \phi_l) \in \mathbb{C}^{(N+1)^2 \times L}$, corresponds to the directions of the microphones and is defined in a similar manner as matrix $\mathbf{Y}(\theta_{sq}, \phi_{sq})$. The mode strength matrix $\mathbf{B}(kr) \in \mathbb{C}^{(N+1)^2 \times (N+1)^2}$ denotes the radial dependence of sound pressure and is mathematically defined as

$$\mathbf{B}(kr) = \mathrm{diag}(b_0(kr), b_1(kr), b_1(kr), \ldots, b_N(kr))$$
$$b_n(kr) = 4\pi i^n \left[j_n(kr) - \frac{j_n'(kr)}{h_n'(kr)} h_n(kr) \right] \tag{3}$$

where $j_n(\cdot)$, $h_n'(\cdot)$ defines the spherical Bessel function of first kind and spherical Hankel function of second kind respectively. $(\cdot)'$represents the derivative. The received sound pressure by the SMA is decomposed in SH domain for the known arrangement of the microphones. Mathematically the sound pressure vector in the SH domain is expressed as

$$\mathcal{P}_{nm}(k) = \mathbf{B}(kr)\mathbf{Y}^H(\theta_{sq}, \phi_{sq})\mathbf{s}(k) + \mathbf{z}_{nm}(k), \quad \forall\, k. \tag{4}$$

where $\mathbf{z}_{nm}(k) = \mathbf{Y}^H(\theta_l, \phi_l)\mathbf{z}(k)$ denotes the noise component after SH decomposition. Subsequently, the SH pressure components are multiplied by the inverse mode strength matrix $\mathbf{B}^{-1}(kr)$ on both sides of the equation (4) to become radial dependency-free.

$$\boldsymbol{\sigma}_{nm}(k) = \mathbf{Y}^H(\theta_{sq}, \phi_{sq})\mathbf{s}(k) + \tilde{\mathbf{z}}_{nm}(k), \forall k \tag{5}$$

where $\tilde{\mathbf{z}}_{nm}(k) = \mathbf{B}^{-1}(kr)\mathbf{z}_{nm}(k)$. Moreover, in order to analyze speech signals that vary over time, the short-time Fourier transform (STFT) of the signal is utilised. Hence the signal model is represented as

$$\boldsymbol{\sigma}_{nm}(t, f) = \underbrace{\mathbf{Y}^H(\theta_{sq}, \phi_{sq})\mathbf{s}(t, f)}_{\boldsymbol{\sigma}^t_{nm}(t,f)} + \underbrace{\tilde{\mathbf{z}}_{nm}(t, f)}_{\boldsymbol{\sigma}^z_{nm}(t,f)} \tag{6}$$

Where $\boldsymbol{\sigma}^t_{nm}(t, f)$ denotes the target component and $\boldsymbol{\sigma}^z_{nm}(t, f)$ denotes the noise. t and f represent the number of time frames and the number of frequency bins, respectively. The proposed work focuses solely on a single intended direct sound source, while all other sources are regarded as direct undesired, contributing to interference. For the sake of ease of understanding $\boldsymbol{\sigma}^t_{nm}(t, f)$ can be written as $\boldsymbol{\sigma}^t_{nm}(t, f) = \boldsymbol{\sigma}^D_{nm}(t, f) + \boldsymbol{\sigma}^U_{nm}(t, f)$. $\boldsymbol{\sigma}^D_{nm}(t, f)$ and $\boldsymbol{\sigma}^U_{nm}(t, f)$ represents the STFT of the desired source and undesired direct interfering sources. Thus the data model in (6) is represented as

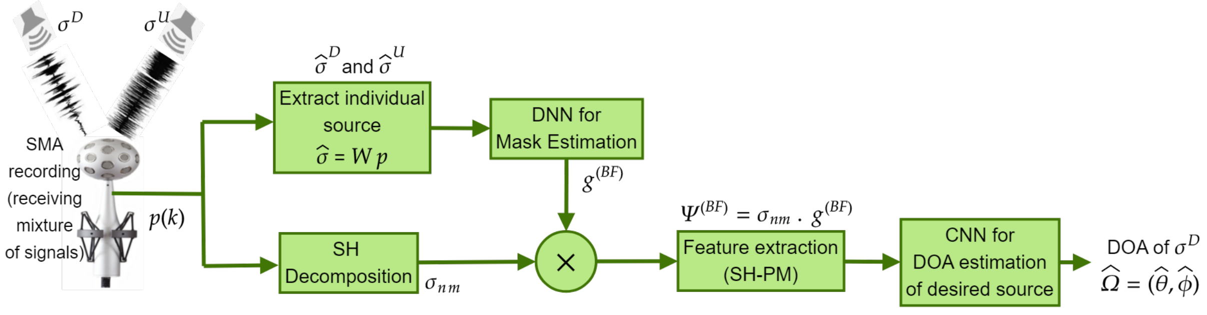

σ^D
σ^U
SMA recording (receiving mixture of signals)
$p(k)$
$\hat{\sigma}^D$ and $\hat{\sigma}^U$
Extract individual source
$\hat{\sigma} = W\,p$
DNN for Mask Estimation
$g^{(BF)}$
SH Decomposition
σ_{nm}
×
$\Psi^{(BF)} = \sigma_{nm}\ .\ g^{(BF)}$
Feature extraction (SH-PM)
CNN for DOA estimation of desired source
DOA of σ^D
$\hat{\Omega} = (\hat{\theta}, \hat{\phi})$

Figure 2: Block diagram of the proposed method. The complete system model and preprocessing of the signal received at the microphone. $\hat{\Omega} \in \{\hat{\theta}, \hat{\phi}\}$ represents the azimuth and elevation directions estimated from the learning model.

$$\boldsymbol{\sigma}_{nm}(t,f) = \boldsymbol{\sigma}^D_{nm}(t,f) + \boldsymbol{\sigma}^U_{nm}(t,f) + \boldsymbol{\sigma}^z_{nm}(t,f) \quad (7)$$

Estimating the DOAs of the desired direct source $\boldsymbol{\sigma}^D_{nm}(t,f)$ in the presence of undesirable direct sources $\boldsymbol{\sigma}^U_{nm}(t,f)$ and ambient noise $\boldsymbol{\sigma}^z_{nm}(t,f)$ is the aim of this paper.

2.2 Source Filtering and Feature Extraction

This section describes the filtering process to remove the undesired sources from the mixture received at the SMA. Subsequently, the features are extracted only for the desired source for its DOA estimation. The SMA signals are the result of combining desired and undesirable source signals with white Gaussian noise in a reverberant environment. To locate the desired sound source, the SH-PM map of the SMA signals is acquired. The SH representation for each microphone channel (η) is given as

$$\Psi^{(\eta)} = \boldsymbol{\sigma}_\eta(t,f), \quad \eta = 1,\ldots,(N+1)^2.\ \forall t,f \quad (8)$$

where $\Psi^{(\eta)} \in \mathbb{R}^{\mathcal{T}\times\mathcal{F}}$. $\boldsymbol{\sigma} = [\boldsymbol{\sigma}_1,\ldots,\boldsymbol{\sigma}_\eta]^T$ are the SH decomposed signals. $\mathcal{T}$ and $\mathcal{F}$ represents the number of time and frequency components in a single frame. The objective is to estimate the elevation and azimuth of the desired source (θ^D, ϕ^D) and set aside the (θ^U, ϕ^U) for Ψ. In this context a binary mask filter is formulated that evicts the features at the time-frequency bin where the undesired source is dominant. Considering that the received signal integrates the two source signals (desired and undesired), separation of individual sources is required for mask estimation. Therefore, to distinguish between the two sources (desired and undesired), an un-mixing matrix is estimated iteratively. Subsequently the estimates of the direct source $\hat{\sigma} = [\hat{\sigma}^D, \hat{\sigma}^U]^T$ are calculated using the observation $\mathcal{P}$ and the corresponding expression is [18]

$$\hat{\sigma} = W\tilde{\mathcal{P}} \quad (9)$$

where W is the unmixing matrix and $\tilde{\mathcal{P}}$ is the white linear transform of $\mathcal{P}$, i.e. $E[\tilde{\mathcal{P}}\tilde{\mathcal{P}}^T] = \mathbf{I}$. The whiten observations are given by $\tilde{\mathcal{P}} = V\Sigma^{-1/2}V^T\mathcal{P}$. Here V denotes the eigen-vectors orthogonal matrix, and Σ denotes the eigen-values diagonal matrix. The eigen value decomposition (EVD) provides the values of V and Σ, i.e. $E[\mathcal{P}\mathcal{P}^T] = V\Sigma V^T$. The components in the unmixing matrix $W = [\mathbf{w}_1,\ldots,\mathbf{w}_Q]^T$ are computed iteratively. Following each iteration, the projection of obtained component with the previous components is obtained and subtracted from the present component. Then the obtained component is normalized. These operations are discussed in detail in [16].

Therefore, the binary filter $g^{(BF)}(t,f)$ can be expressed as

$$g^{(BF)}(t,f) = \begin{cases} 1 & \text{if } \mathcal{R}^{(BF)} \geq \nu_{th} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where $\mathcal{R}^{(BF)}$ is the ratio to obtain the mask and is optimized by the DNN, and $\nu_{th} \in [0,1]$ is the threshold. In this paper, a DNN-based binary filter is designed for this purpose to obtain real-time masks. The DNN framework for mask estimation and the loss function to optimize $\mathcal{R}^{(BF)}$ is discussed in section 3.1. So, the features of desired sources are obtained by suppressing the predominate feature components of the interfering direct source for training the CNN model. These characteristics are denoted as follows

$$\Psi_{BF}^{(\eta)} = \boldsymbol{\sigma}_{\eta}(t,f)g^{(BF)}(t,f); \quad \eta = 1,\ldots,(N+1)^2 \tag{11}$$

Figure 2 illustrates the complete end-to-end block diagram of the proposed model for source separation, extracting the features, and learning architecture.

3. LEARNING FRAMEWORK AND DATASET GENERATION

This section addresses the learning model developed for the estimation of the mask for filtering out the undesired source signal and another learning model for the DOA estimation using CNN architecture. Figure 3 depicts the network design for both DNN and CNN. Also the experimental conditions for the simulated as well as real-time dataset generation is discussed herein.

3.1 Learning Framework for Mask Estimation

For obtaining the binary mask mentioned in equation (10), a DNN is trained for each channel. The input to the DNN is the received SMA signal $|\sigma_{nm}(:,:)|$. It has a feed-forward output layer with sigmoid activation that calculates a ratio mask $\mathcal{R}^{(BF)} \in [0,1]$ and three bidirectional long short-term memory (BLSTM) layers consisting of 1200 neurons per layer. To train the DNN, the mean-squared error (MSE) loss function ($\mathcal{L}$) is minimized for each channel. The MSE is taken between the desired and estimated (received microphone signal multiplied with ratio mask) signals and is defined as

$$\mathcal{L} = \sum_{f=1}^{F}\sum_{t=1}^{T} \frac{(|\hat{\sigma}_{nm}^{D}(t,f)| - |\sigma_{nm}(t,f)|.\mathcal{R}^{(BF)})^2}{T.F} \tag{12}$$

where T and F are the total number of time and frequency frames, and $\hat{\sigma}_{nm}^{D}(t,f)$ is the direct sound of the desired source signal, obtained from equation (9). The binary mask ($g^{(BF)}$) is finally obtained as given in equation (10).

3.2 Learning Framework for DOA Estimation

After breaking down the signal into its component SH domains, the proposed work develops an SH-CNN model for DOA estimation. The DOA is estimated only for the desired sources after eliminating the undesired sources by filtering using the masks obtained from the previous section. The convolutional network improves the classification of the data by automatically identifying the key patterns in the input features, resulting in robust learning. The input characteristics used to train the CNN model are the SH magnitude and phase coefficients. The appropriate azimuth and elevation classes are given to the input SH features. Following batch normalisation and max-pooling, the CNN network consists of three convolutional layers and two dense layer. With the exception of the output layer, each layer's activation function is a rectified linear unit (ReLU). The maximum probability determines the azimuth or elevation estimations, and the likelihoods corresponding to each class are obtained in the output layer using the softmax activation function. Furthermore, the dense layer's dropout is used to prevent over-fitting. In the work that is being presented, a dropout of 0.20 is used in the dense layer. The output layer gives the source's azimuth and elevation DOA estimates.

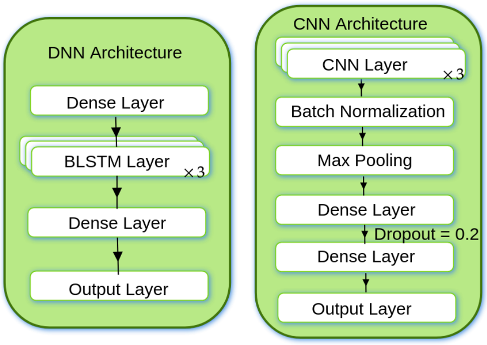
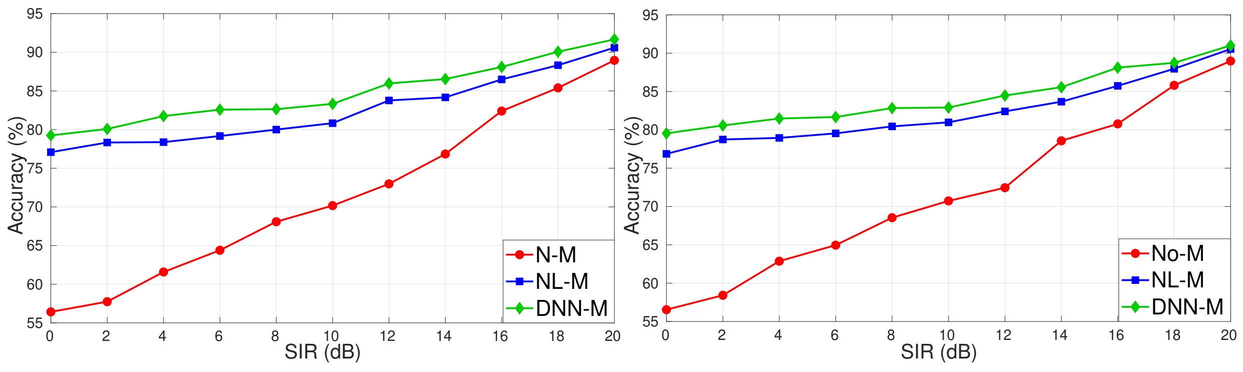


Figure 3: The DNN model for mask generation and the CNN model for learning and estimating DOA of desired sources.

3.3 Experimental Conditions and Dataset Generation

Extensive simulation is carried out for generating data using the SMA impulse response (SMIR) generator. The simulation room is 5 m by 6 m by 7 m in size, with a variance of ±2 m in each dimension. The reverberation time (RT$_{60}$) is randomly selected from $(0.2-1)s$. The training data also includes random white noise with an SNR $\in [0-20]$dB, in addition to the reverberation. The sources and SMA are positioned with a minimum angular separa-



(a)

(b)

Figure 4: Illustration of accuracy (ACU) (in dB) by varing SIR, for N-M, NL-M, and DNN-M localization methods for (a) Azimuth estimation and (b) Elevation estimation.

Table 1: Performance analysis RMSE [°] for DOA estimation of desired source. $(\Omega)^D \in \{\theta^D, \phi^D\}$ is the desired source and $\Omega^U \in \{\theta^U, \phi^U\}$ is the undesired source

	$\phi^D = 60°,\ \theta = 30°$					$\theta^D = 60°,\ \phi = 30°$		
ϕ^U	SIR	N-M	NL-M	DNN-M	θ^U	SIR	N-M	NL-M	DNN-M
45°	5	4.6	3.8	3.2	45°	5	4.6	3.4	3.1
	10	4.3	3.5	3.2		10	4.2	3.4	3.2
	15	4.2	3.1	2.7		15	4.2	2.9	2.7
	20	4.1	2.8	2.6		20	4.1	2.7	2.5
150°	5	2.9	1.4	1.3	150°	5	3.1	1.3	1.2
	10	2.8	1.3	1.1		10	3.1	1.2	1.1
	15	2.6	1.2	1.0		15	2.6	1.1	1.1
	20	2.3	1.1	0.8		20	2.3	1.1	0.8

tion of 5° and a distance between them that varies from 2 to 5 m. The LIBRISPEECH [19] and FSDNOISY18K [20] libraries are used to choose the desirable and undesirable sources, respectively. Signal-to-interference ratio (SIR) $\in [0-20]$ dB is used to integrate the two directional source signals. For the training data generator, source signals that are 1s long and sampled at 16 KHz are taken into consideration. An Eigenmike [21] is used to acquire sound scenes. The Eigenmike is made up of a 32 flush-mounted microphone placed over an order 4-hybrid rigid sphere that measures 4.2 cm in diameter. STFT is used to examine the received signals in the frequency domain with a Hanning window that has a 512 length and a 50% overlapping.

4. PERFORMANCE ANALYSIS

The performance of the proposed DNN masking (DNN-M) method is discussed in this section. The recordings from the SMA are taken and the DNN-based mask is applied to it for separating the undesired source and then the filtered signal is given as input to the CNN model to estimate the DOAs. The performance is compared with [16] where mask estimation is not done by learning method i.e. no learning masking (NL-M). Also the performance is compared with the case if no masking (N-M) is applied and the recording containing the mixture of sources is directly given as CNN input for DOA estimation.

4.1 Numerical Analysis

Extensive simulations are carried out for analysing the performances of various methods. Root mean square error (RMSE) and accuracy (ACU) are considered for the comparison of the performances. The RMSE and ACU are expressed as

$$\text{ACU}(\%) = \frac{\hat{\delta}}{\delta} \times 100$$

$$\text{RMSE}_{\Omega} = \sqrt{\frac{1}{\delta}\sum_{i=1}^{\delta}(\Omega_i - \hat{\Omega}_i)^2} \tag{13}$$

where δ are total number of DOAs and $\hat{\delta}$ are accurately estimated DOAs. $\Omega \in \{\theta, \phi\}$ are given elevation or azimuth angles and $\hat{\Omega} \in \{\hat{\theta}, \hat{\phi}\}$ are estimated elevation or azimuth angles. The method proposed, DNN-M, is

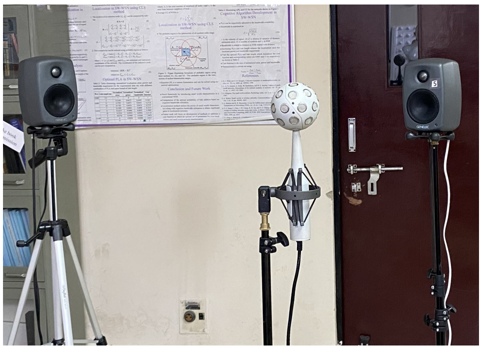


Figure 5: Experimental lab set up for acoustic data acquisition

Table 2: Results for the experimental setup at IIT Kanpur of DOA estimation for the three test cases

Test cases	Ω^D	Ω^U	Metric	N-M	NL-M	DNN-M
C-I	$45°, 60°$	$90°, 150°$	ACU	50.8	75.0	78.6
			RMSE	5.4	4.1	3.5
C-II	$30°, 290°$	$110°, 55°$	ACU	51.2	76.6	79.9
			RMSE	5.4	3.9	3.3
C-III	$120°, 70°$	$120°, 25°$	ACU	50.5	74.8	78.0
			RMSE	5.5	4.2	3.6

compared with the case which filters the undesired source without learning-based masking, i.e. NL-M [16] and the case when DOA is estimated even without suppressing the undesired source, i.e N-M. The evaluation of the performance for all the methods is done at various signal to interference ratio (SIR) between $[0 - 20]$ dB for azimuth as well as elevation model as shown in Fig. 4. Also, RMSE values showing the performance of these methods at different test cases for elevation and azimuth angles is shown in Table 1. The figures and table show that the proposed method performs best in all the cases. The performance is much improved than in the N-M case. Moreover, when compared with the NL-M case, the DNN-M case is better because the DNN mask is estimated using the regression approach utilising the optimized mask values to suppress the undesired source. The DNN-optimized mask is better than the conventional mask, which is not well-optimized. Therefore, achieved results are much improved and hence motivating to use for various applications.

4.2 Experimental Results

The experimental analysis for the real-time recordings are also carried out in the lab environment. Figure 5 shows the set-up arranged in the MiPS lab at IIT Kanpur for the experimental verification of the proposed method. Three test cases are taken, C-I, C-II, and C-III, mentioned in Table 2. For the direct desirable and direct undesired interfering sources, respectively, audio files of 5 s duration are selected from the LIBRISPEECH [19] and FSD-NOISY18K [20] libraries. Table 2 contains the numerical results for the ACU and RMSE. When compared to the N-M approach, the proposed method exhibits a considerable reduction in RMSE and significant improvement in ACU for the real-time tests. Also there is an improvement in the DNN-M approach as compared to the NL-M method. These findings indicate that the suggested method can effectively distinguish between the interference and the source, which is necessary for accurate DOA estimation in both azimuth and elevation directions using SH decomposition.

5. CONCLUSION

This research focuses on a learning-based approach to acoustic source localisation in environments where directional interference from an unwanted source is present. SH decomposition is used to get the SH-PM features that match the SMA recordings. In order to generate training datasets, DNN is used to generate a binary mask based on the SH characteristics. Both simulated and real-world test scenarios are used to assess the effectiveness of the suggested approach. The accuracy in low SIR situations is much enhanced by the proposed technique. The proposed method is useful for a wide range of purposes, including but not limited to voice enhancement and localization, teleconferencing, augmented and virtual reality, and more. Future work will expand on this to account for cases with numerous strongly interfering directional sources.

6. ACKNOWLEDGMENTS

This work is supported by the SERB-DST under project no. SERB/EE/2021410.

7. REFERENCES

[1] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust doa estimation of multiple speech sources," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2287–2291, 2014.

[2] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans-*

actions on Acoustics, Speech, and Signal Processing, vol. 24, no. 4, pp. 320–327, 1976.

[3] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2016.

[4] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband music: Opportunities and challenges for multiple source localization," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 18–21, 2007.

[5] 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 Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.

[6] V. Varanasi, H. Gupta, and R. M. Hegde, "A deep learning framework for robust doa estimation using spherical harmonic decomposition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1248–1259, 2020.

[7] P. Dwivedi, R. P. Gohil, G. Routray, V. Varanasiy, and R. M. Hegde, "Joint doa estimation in spherical harmonics domain using low complexity cnn," in *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5, 2022.

[8] P. Dwivedi, G. Routray, and R. M. Hegde, "Doa estimation using multiclass-svm in spherical harmonics domain," in *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5, 2022.

[9] P. Dwivedi, S. B. Hazare, G. Routray, and R. M. Hegde, "Long-term temporal audio source localization using sh-crnn," in *2023 National Conference on Communications (NCC)*, 2023.

[10] P. Dwivedi, G. Routray, and R. M. Hegde, "Hybrid sh-cnn-mp approach for super resolution doa estimation," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pp. 96–100, 2022.

[11] P. Dwivedi, G. Routray, and R. M. Hegde, "Far-field source localization in spherical harmonics domain using acoustic intensity vector," in *24th International Congress on Acoustics (ICA)*, 2022.

[12] P. Dwivedi, G. Routray, and R. M. Hegde, "Learning based method for near field acoustic range estimation in spherical harmonics domain using intensity vectors," *Pattern Recognition Letters*, vol. 165, pp. 17–24, 2023.

[13] S. Sivasankaran, E. Vincent, and D. Fohr, "Keyword based speaker localization: Localizing a target speaker in a multi-speaker environment," pp. 2703–2707, 09 2018.

[14] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[15] W. Mack, U. Bharadwaj, S. Chakrabarty, and E. A. P. Habets, "Signal-aware broadband doa estimation using attention mechanisms," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4930–4934, 2020.

[16] P. Dwivedi, G. Routray, and R. M. Hegde, "Spherical harmonics domain-based approach for source localization in presence of directional interference," *JASA Express Letters*, vol. 2, no. 11, p. 114802, 2022.

[17] B. Rafaely, *Fundamentals of spherical array processing*, vol. 8. Springer, 2015.

[18] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[20] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25, 2019.

[21] "The eigenmike microphone array, [online]. available: http://www.mhacoustics.com/," 2013.