IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 3, June, 2024, pp. 1429-1439
Journal homepage: http://iieta.org/journals/ts

Blind Sound Source Separation by Combining the Convolutional Neural Network and Degree Separator

Swapnil G. Mali*, Shrinivas P. Mahajan

Electronics and Telecommunication Department, COEP Technological University, Pune 411005, India

Corresponding Author Email: sgm15.extc@coeptech.ac.in

https://doi.org/10.18280/ts.410331

Received: 30 March 2023
Revised: 27 October 2023
Accepted: 15 November 2023
Available online: 26 June 2024

Keywords:
artificial neural network, blind sound source separation, convolutional neural network-direction of arrival deep learning, degree separator, hybrid algorithms, microphone arrays, soft computing

ABSTRACT

The objective of blind sound source separation is to separate and extract distinct audio sources from a mixture of audio signals with little to no prior information about the mixing process an innovative two-stage approach is presented in this research paper that addresses the challenge of blind sound source mixing within multichannel sound recordings. The paper proposes a two-stage method that combines a Convolutional Neural Network (CNN) and a degree separator to solve the problem of blind sound source mixing in a multichannel sound recording. The first stage uses CNN to estimate each sound source's Direction of Arrival (DOA) in each time frame. The second stage consists of a degree separator that separates the target source from multiple sources by converting the signal from convolutional to the linear domain. The effectiveness of the proposed method is extensively evaluated using a range of sound sources, including recordings of real-world audio databases created using simulated and actual room impulse responses The estimated DOA of each source is compared against the ground truth trajectory of each source within the complex, multi-sourced environment. The degree separator evaluation is based on Blind Source Separation (BSS) evaluation criteria compared to Fast Independent Component Analysis (FICA). Source separation performance is evaluated using multiple sound sources in simulated and room impulse response recording. The proposed method is evaluated by separation quality parameters such as the image-to-spatial distortion ratio (ISR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR). The proposed method is evaluated using both simulated sound sources and real room impulse response recordings. This research presents a powerful solution for estimating DOA of multiple sound sources and effectively separating them in multichannel sound recordings. Based on comprehensive evaluations performed on stationary and moving source in simulated and actual room condition. The proposed method surpasses conventional BSS approaches regarding separation quality by combining CNN-DOA with a degree separator.

1. INTRODUCTION

Human beings can extract a source of interest from an audio mix in real-time by using sensed information from the ear. Source separation removes a target speech or sound from a particular source in-room environment or open space. Sound source separation is a challenging and emerging research area. Researchers try to develop real-life applications such as robot audition, assisting listening devices, meeting transcription systems, Automatic Speech Recognition (ASR), 3D sound effects, and many other applications [1]. When no prior or little information about the captured sources is available, the process is called Blind source separation (BSS) [2]. The BSS problem involves reconstructing a signal from a mixed signal or a set of mixed signals. Many different source separation systems are available, including multichannel, monaural, and room source separation. Independent Component Analysis (ICA) is a traditional BSS technique [3, 4]. ICA creates a contrast function to demix signals using maximizing non-Gaussianity and minimization of mutual information. ICA fails to separate mixed signals in a reverberant room environment. In the frequency domain ICA, it faces two problems: the first problem is the permutation of each source; the second is the scaling problem of each source signal [4, 5]. Researchers have proposed various methods to solve these two problems in ICA. The Time Difference of Arrival (TDOA) method is used to solve the permutation problem of ICA [6]. In TDOA, if the source frequency exceeds the spatial aliasing limit, source location estimation becomes ambiguous. Therefore, TDOA in ICA is not valid for the high-frequency source signal. Beamforming with ICA techniques can also be applied to BSS to improve the separation performance [4]. Resent beamformer adaptively estimates noise characteristics and the sidelobe canceller [7]. However, it requires many microphones in a physically more extensive linear array to form a narrow beam to separate closely spaced sources. Some beamforming cases need a complex array and a denser sensor arrangement on spherical geometry, which is impractical in real-life applications [8, 9].

Nonnegative Matrix Factorization (NMF) helps separate

sound sources in single and multichannel mixtures [10, 11]. The standard NMF technique is more suitable for single-channel separation. In NMF, the algorithm converts a mixed-signal spectrogram into a product of two nonnegative matrices. One matrix is a basis vector representing source information, and the other one is a basis vector activity matrix indicating the time-varying gain for each basis vector [12, 13]. All channels' magnitude or power spectrograms are stacked into nonnegative tensors in the multichannel NMF model. An STFT coefficient is a complex-valued realization of a zero-mean Gaussian random variable [14]. An NMF-based separation is more useful when the environment is weakly guided, and the information is limited. NMF fails to account for inter-channel phase difference in its spectra-temporal magnitude model. NMF with a fixed number of NMF components per source also gives less separation accuracy [11].

A BSS system with high localization accuracy and adaptability in dynamic acoustic scenarios with multiple source conditions is a challenging task. The primary objective of the research work is to create and analyze a novel method for the separation of mixed audio sources in a blind source separation scenario. The separation of sound sources is accomplished by combining the strength of Convolutional Neural Networks (CNNs) for feature extraction with the Degree Separator technique. Generally, CNN is used in image classification in two-dimensional data, and this paper introduces CNN in speech processing as a preprocessing stage to the existing BSS problem. Here, the cross-correlation between inter microphones and a particular source in the STFT frame is utilized for training the CNN-DOA framework before the separation stage processing [15]. A Convolutional Neural Network (CNN) is used to estimate the Direction of Arrival (DOA) of a sound source in an audio signal by analyzing its spectrogram or other time-frequency representations. The CNN is trained on labeled data containing audio recordings with known DOA information. During inference, the CNN applies a set of learned filters to convolve over the input spectrogram, extracting relevant spatial features indicative of the source's DOA. These features are then processed through additional layers to predict the DOA angle.

This paper extends the work on DOA estimation of multiple speakers using a CNN-based approach. The training of the system is carried out in diverse acoustic scenarios and multi-source -conditions to make it a more robust in-room environment. The proposed method uses a degree separator for source content separation and DOA estimation of the sources using CNN. The term "degree separator" is an algorithmic procedure that utilizes information of the impulse response at each source location to separate mixed audio source signals. Separation is accomplished by iteratively modifying coefficients in the linear equations to optimize a cost function after transforming the mixing signal from a convolutional domain to a linear domain. The separation quality of sound is assessed using evaluation parameters for BSS such as SNR, SIR, SDR, and STOI [16]. The combination of a CNN and degree separator leverages the strengths of deep learning for feature extraction and the mathematical optimization. The proposed method produces a significantly better separation quality than traditional BSS methods.

The remaining paper is organized as follows: Section 2 discusses the experimental setup and methods for creating databases. Section 3 describes the proposed methodology. The results of CNN -DOA and the degree separator evaluation are given in Section 4. Section 5 discusses the conclusion and the scope of future work.

2. EXPERIMENTAL SETUP AND DATABASE CREATION

This section discusses the source mixing model for representing the signal and the experimental setup to create a database using simulated and recorded Room Impulse Response (RIR). To develop a BSS system, we need to understand the mixing process in the anechoic and typical room environments. Eq. (1) denotes linear mixing in an anechoic room. Here upper case denotes matrices, and t denotes the time index. Consider the condition of multiple sources in an anechoic room environment, and the signal is recorded using a microphone array. Multiple source signals are mixed linearly by Eq. (1).

$$\mathrm{X_m} = \mathrm{AS_k} \tag{1}$$

where, k =1, 2, …. K are the various sources, the number of microphone m=1,2,3, …. M, the number of samples of each source signal n=1,2, … N, and $\mathrm{S_k}$ is the Source signal matrix with K * N, A is the Mixing matrix with dimensions M *K and $\mathrm{X_m}$ is the matrix of mixed-signals with dimension M *N. The convolutive mixing in-room environment [17] is presented by Eq. (2) below:

$$x_{mix}(t) = \sum_{p=1}^{P} \sum_{\tau} S_p(t-\tau) h_{pmt}(\tau) \tag{2}$$

Here, the microphone ranges from m=1.....M and $x_m(t)$ is the mixed-signal of length p=1…. P, source signals $S_p(t)$ are sampled at the discrete-time. If sources are moving, then the room impulse response $h_{pmt}(\tau)$ has time-varying mixing properties. The aim is to estimate the source signal $S_p(t)$ with estimated $h_{pmt}(\tau)$ and a known mixed-signal $x_m(t)$. Linear mixing in Eq. (1) is a simplified model that assumes instantaneous mixing of source signals in anechoic signal. In contrast, convolutive mixing in Eq. (2) is the complex interactions of sound in a room, which includes reflections and delays due to room impulse responses. This model is used for creation of mixed audio database for training and testing of proposed CNN-DOA method and implementation of degree separator method.

In this proposed work, the BSS method involves estimating the DOA and separating the source without source information. DOA is the direction from which the sound is emitted towards the microphone. In this case, the DOA of the source is not available, i.e., the source location is unknown; only a database of the room impulse response in different directions in the room is available. Estimating accurate DOA becomes more challenging for multiple sources in a room environment. RIRs are essential for creating a database that will be used to train and test the CNN-DOA system. RIRs offer realistic representations of acoustic environments, including echoes and reflections in the room, adding variability that enables the model to be generalized to other room setups. Mix audio signal database is created by convolving source signals with various RIR. The experimental setup consists of two types of RIR responses: Simulated room impulse responses [18] and the other is RIR database from Bar-Ilan University [15, 18, 19]. Different acoustic conditions are created in a simulated

environment with different parameters, as shown below in Table 1, to create different room conditions. Variation in the locations of source arrays in the room is introduced to develop robustness in the acoustic environment during the training of the model. The simulated RIR database Image-based method is used for simulating a small, acoustic room impulse response with a wide range of room parameters while maintaining accurate control of the experimental conditions. Users can set various parameters in this environment like sampling frequency, the position of the microphone array, the distance between the microphones, the type of microphones, and the location of the source to ULA. Reflection coefficient, reverberation time, and location parameters can be set to generate RIR for a particular location. Table 1 shows the various parameters for the simulated acoustic environment for database creation.

RIR database from Bar-Ilan University database: In this research, the second type of RIR database is a Multichannel RIR database from Bar-Ilan University [18, 19]. Impulse responses are measured in the Speech & Acoustic Lab of the Faculty of Engineering at Bar-Ilan University. Details of parameters used by them are specified in Table 2. We have used RT60 for experimentation with the acoustics environment for RIR of different positions. It was recorded at a distance of 1m and 2m from the center of the ULA. Seven source positions were considered, along with a semicircular grid covering the whole angular range of 0° to 180° with a step size of 30°. The inter microphone distance for eight microphones ULA was 0.05 m. This RIR database consisted of eight microphones RIR with different locations in the room environment. Table 2 shows various parameters from the Bar-Ilan University database.

Two types of databases are created: one with simulated RIR, and the other one is the RIR database from Bar-Ilan University. The RIR recoded signal is convolved with a speech signal from the LIBRI database, and a WGN is created using Audacity with different SNR levels of 5 dB, 15dB, and 25 dB to create a signal for training and testing of CNN-DOA and the degree separator. The mixed audio database is created by convolving the source signal with the simulated RIRs or Bar-Ilan University RIRs. A single source signal is created by one source signal convolving with one RIR of the corresponding location; the resultant convolved signal is a single source signal in a given room environment. A Mixed signal of two sources is created by adding two convolved signals. The first convolved signal is created by one source convolution with RIR of one location in the room, and the second signal is created by another source convolved with RIR of a different location of the same room. Using the same mentioned technique, a three-source database is generated using three source signals and three RIRs.

Table 1. Parameters in a simulated acoustic environment for database creation

Room Size	4m × 6m ×3m OR 5m × 7m × 3m, OR 5m × 6m × 4m
Reverberation Time: RT (60)	0.16s, 0.36s and 0.61s
ULA and Microphone Distance	8 Microphone ULA with different Inter microphone distance
DOA Resolution	30 ° (from 0°to 180°)
Source - Array Distance	1m and 2 m
Sound Source Signal	The speech signal from LIBRI and WGN was created using Audacity with different SNR levels of 5 dB,15 dB, and 25 dB for training and testing

Table 2. Parameters of the acoustic environment in an RIR database created at the Bar-Ilan University

Room Size	6m × 6m × 2.4m
Reverberation Time: RT (60)	0.16s, 0.36s and 0.61s
ULA and Microphone Distance	8 Microphone ULA with different inter microphone distances
DOA Resolution	15° (from 0°to 180°)
Source - Array Distance	1m and 2 m
Sound Source Signal	The speech signal from LIBRI and WGN was created using Audacity with different SNR levels of 5 dB,15 dB, and 25 dB for training and testing

3. PROPOSED METHODOLOGY

The proposed blind sound source model is shown in Figure 1. The model is based on the DOA estimation of the sources using CNN and source content separation using a degree separator. The first stage involves estimating sound signals in each time frame using CNN, and the second stage consists of a degree separator that separates the target source from a mixture of multiple sources using a convolutive form to the linear conversion process. The DOA estimation using a CNN (CNN – DOA) methodology estimates the DOAs of many concurrently active sources in simulated and real-world situations. CNN- DOA estimation consists of possible N-classes. A set of possible DOA values are $\Theta = \{\theta 1, \theta 2, \ldots, \theta I\}$. Possible source locations can be 0°, 30° up to 180° with a special resolution of 30°. Here seven different classes for each DOA are considered for experimentation with the assumption that there is no overlap between source locations in multiple source scenarios.

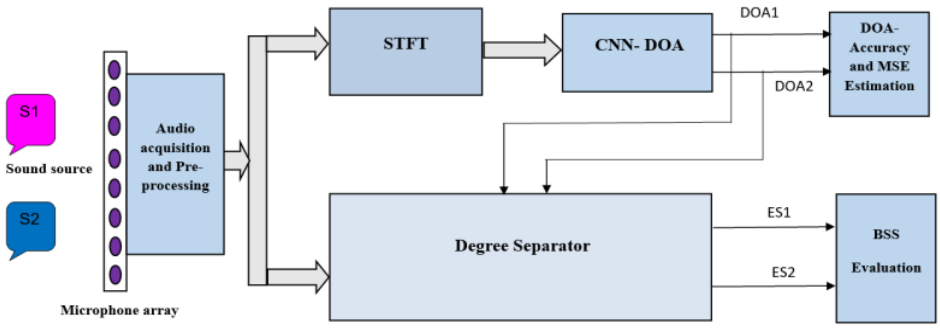

Figure 1. The proposed system for blind sound source separation

The maximum sources in the room are three with two cases: one with all three being static and the other with one or two moving sources. The remaining sources are considered static sources. The goal of the CNN-DOA method is to use mixed-signal frames to estimate the DOA of many speakers with static and moving sources. The features provided for training and testing the model are Short-Time Fourier transform (STFT) and mixed-signals recorded in multiple source position scenarios. The DOA of multiple sources is estimated based on blocks of STFT frames of the observed mixed signal. The STFT block length depends on dynamic or static multiple sources in a simulated and actual room environment. CNN-DOA is a supervised learning system that includes training and testing phases using audio STFT frames as input images. This method is trained with an STFT feature data set corresponding to a specific mixed-signal recorded with the known DOA of each source. This true DOA class has a corresponding label in each STFT frame. In the test phase, we first estimate the DOA class of each STFT frame and then estimate a class of STFT block length by averaging the probabilities of all STFT frames. The DOA estimates are then computed by identifying the DOA classes with the highest probability. We assume that the number of sources actively participating in the scenario is known to us. Degree separators consist of estimation of source signals using knowledge of the source location room environment. For the purposes of this experiment, in the simulated room environment, two stationary sources, S1 and S2, are considered active sources, and a mixed signal is recorded using 8 linear microphone arrays. We have assumed two sound sources, S1 and S2, at a specific location (here, the location of each source is estimated by CNN -DOA). The mixture at mic one is mathematically represented by the following Eq. (3):

$$Xmix1 = h11 * S1 + h21 * S2 \quad (3)$$

where, Xm1 is the mixture recorded at mic one, S1is the first source of N samples, and h11 is the RIR between source S1 and mic 2. S2 is the second source of the N sample, and h21 is the RIR between source S2 and mic M2. The feature used for CNN -DOA is STFT on the audio signal. Using STFT, one can transform an audio signal into an image. STFT consists of two components, namely, the magnitude component and the phase component. Here the extracted STFT image is created for each time frame using a Hanning window of Nf samples. The Fast Fourier transform used in STFT is Nf, which leads to an STFT image size of ((Nf /2) + 1) x k for an audio signal. Where k is the number of frames in the audio signal, Nf = 512, and the size of the STFT image is 257 x k. We extract the magnitude and phase components of each STFT image. Now, the input audio signal Sm(k,b) can be represented in magnitude and phase parts as follows:

$$Sm(k,b) = Am(k,b) * e^{j\phi_m(k,b)} \quad (4)$$

where, A = magnitude component, ϕ = phase component, m = number of microphones, k = time frame and b = frequency bin. After using STFT images as magnitude and phase components separately in CNN DOA experimentation, it is observed that phase components are more essential for source localization compared to magnitude components. That magnitude component has a relatively less significant role in the localization of sound sources. The size of the STFT phase component of each microphone is 257 × k as one audio mixture consists of m versions of the same signal in m microphones. In our case, m= 8, so for one audio mixture, the size of a 3-D matrix is 257 × k × 8. This 3 -D matrix belongs to each DOA class that is provided for training. Here k input images with size 8 × 257 are provided for training of CNN-DOA. CNN – DOA was trained based on location-dependent sources and mics phase variation embedded in an input STFT image of size 8 × 257.

3.1 The DOA estimation using CNN (CNN -DOA)

CNN is the most popular algorithm used widely for image classification, object detection, natural language processing, and speaker identification. CNN is used to identify and separate the various features of an image input [20-22]. An STFT phase map is provided as an input image to CNN in this model. In general, CNN mainly consists of different layers such as the input layer, convolutional layer, pooling layer, fully connected layer, softmax layer and output layer. A Convolutional Neural Network (CNN) is used to estimate the Direction of Arrival (DOA) of a sound source in an audio signal by analyzing its spectrogram or other time-frequency representations. The CNN is trained on labeled data containing audio recordings with known DOA information. The CNN-based DOA estimation model is capable of localizing sound sources in various applications, such as microphone arrays, robotics, or acoustic scene analysis. The architecture of CNN is presented in Figure 2.

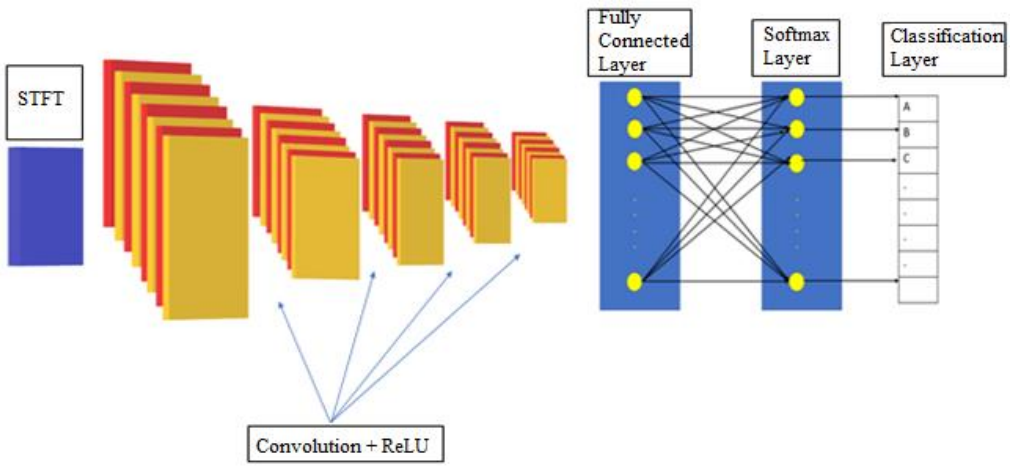

Figure 2. The architecture of CNN – DOA

1. Input Image: The first layer is the image with dimensions 8 × 257. This image is taken from the phase component 3-D matrix size 257 × k × 8.

Here K images of size 8 × 257 are selected for one class from the input. These images serve as the starting point for further processing.

2. Convolutional Layers (Layers 2-5): The CNN employs convolutional layers, starting from the second layer up to the fifth layer with 32 or 64 convolution filters, and the filter has sizes like 2 × 2 or 3 × 3. In this case, the input image is of a small size, i.e., 8 × 257; hence, a filter size of 5 × 5 or more is not practically possible. So, to enhance the accuracy of the CNN model, deeper layers of the network are created. These layers use either 32 or 64 convolution filters. The filters used are in various sizes, 2 × 2 or 3 × 3 instead of 5 × 5, to employ a deeper network to enhance the model's accuracy.

3. Activation Function (ReLU): The Rectified Linear Unit (ReLU) is the activation function used. ReLU is more reliable and accelerates convergence than the sigmoid and tanh functions.

4. Fully Connected Layer (Layer 6): The sixth layer is fully connected. It is used to learn non-linear features as represented by the output of the convolutional layer. The output of the last convolutional layer is flattened and fed into the fully connected layer.

5. Softmax Classification Layer (Last Layer): The sixth layer is the fully connected layer. The last layer is a softmax classification layer that tackles multiclass classification issues. It's a layer with N potential classes, which depend on N in various combinations depending on the location of the different sources in the room.

6. The CNN architecture uses the following hyper parameters:

Learning Rate: 0.001
Loss Function: Cross-Entropy
Optimizer: Adam
Number of Epochs: 20 to 50
Batch Size: 64 to 128
Regularization: Dropout (0.5 dropout rate)

These hyper parameters regulate the model's training process and can significantly impact model performance.

This CNN architecture is designed for audio source classification tasks, aiming to correctly classify sound sources in various room environments with many sources. This CNN architecture uses convolution layers, ReLU activation, and small filter sizes to extract relevant special features from the input image.

3.2 Degree separator

The degree separator is a novel approach implemented using a synthesis model and an estimation model concept of mixed-signal. Figure 3 shows how a mixed-signal creates an input recording of two sources from two directions using the convolutive mixing of two sources.

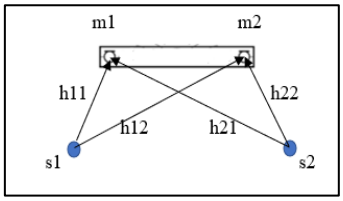

Figure 3. Actual mixing of two sound sources in a room environment

Now consider the case of two sources in convolutive mixing (Room recording) with two mics given by the following equation.

where, $h_{ij} = room\ impulse\ response\ of\ source$ i $to\ mic$ j,

$$x^1_{mix}(n) = x^1_1(n) + x^1_2(n) \tag{5}$$

$x^1_1 = mix\ signal\ at\ mic\ 1\ due\ to\ source\ 1$, and $x^1_2 = mix\ signal\ at\ mic\ 1\ due\ to\ source\ 2$

$$x^1_{mix}(n) = h_{11} * s_1 + h_{21} * s_2 \tag{6}$$

where, $x^1_1(n) = h_{11}(n) * s_1(n)\ and\ x^1_2(n) = h_{21}(n) * s_2(n)$.

Consider the synthesis model, i.e., a simplified mathematical model for synthesizing the mixed-signal in-room environment with two sources and four different room impulse responses with order P.

Consider the Synthesis model for the synthesis of x mix signals at Mic1 and Mic 2 for n=0.

$$\begin{bmatrix} x^1_{mix}(0) \\ x^2_{mix}(0) \end{bmatrix} = \begin{bmatrix} h_{11}(0) h_{21}(0) \\ h_{12}(0) h_{22}(0) \end{bmatrix} \begin{bmatrix} s_1(0) \\ s_2(0) \end{bmatrix}. \tag{7}$$

where, $x^1_{mix}\ and\ x^2_{mix}$ are the mixed-signals received at Mic 1 and Mic 2, respectively.

Similarly for n=1

$$\begin{bmatrix} x^1_{mix}(1) \\ x^2_{mix}(1) \end{bmatrix} = \begin{bmatrix} h_{11}(0) h_{21}(0) \\ h_{12}(0) h_{22}(0) \end{bmatrix} \begin{bmatrix} s_1(1) \\ s_2(1) \end{bmatrix} + \begin{bmatrix} h_{11}(1) h_{21}(1) \\ h_{12}(1) h_{22}(1) \end{bmatrix} \begin{bmatrix} s_1(0) \\ s_2(0) \end{bmatrix} \tag{8}$$

where, all impulse matrices are of size 2*2 with Notation H0, H1 ……. Hp. In the degree separator (Separating System), convolved mixing is converted into linear mixing, and then the mixed-signal samples are converted into respective source signal samples. The degree Separating System is given in the following steps, and the estimated value in the first step is used in the next step, as shown below:

Step 1: Consider source at n=0 using the following equation from Eq. (7):

$$\begin{aligned} x^1_{mix}(0) &= h_{11}(0).* s_1(0) + h_{21}(0).* s_2(0) \\ x^2_{mix}(0) &= h_{12}(0).* s_1(0) + h_{22}(0).* s_2(0) \end{aligned} \tag{9}$$

Consider sources S1 and S2 at n=1 using the following equation from Eq. (8):

$$\begin{aligned} x^1_{mix}(1) &= h_{11}(0).* s_1(1) + h_{21}(0).* s_2(1) \\ &+ h_{11}(1).* s_1(0) + h_{21}(1).* s_2(0) \\ x^2_{mix}(1) &= h_{12}(0).* s_1(1) + h_{22}(0).* s_2(1) + \\ & h_{11}(1).* s_1(0) + h_{21}(1).* s_2(0) \end{aligned} \tag{10}$$

And so on for sources, S1 and S2 at n=0,1, 2, …. P. The two sources CNN-DOA model estimates the location of both sources S1 and S2, and an appropriate H matrix based on the CNN-DOA classification from the RIR database is selected $h_{11}(0), h_{21}(0), h_{12}(0), h_{22}(0)$, are known parameters along with the mixed signals, $x^1_{mix}(0)\ and\ x^2_{mix}(0)$.So, the problem is to estimate $s'_1(0)\ and\ s'_2(0)$, i.e., a source at n=0 using the following equation:

$$s'_1(0) = h'_{11}(0).* x^1_{mix}(0) + h'_{21}(0).* x^2_{mix}(0) \tag{11}$$

$$s'_2(0) = h'_{12}(0).* x^1_{mix}(0) + h'_{22}(0).* x^2_{mix}(0)$$

Here the aim is to estimate all samples of S1 and S2 without any prior knowledge about these sources. The error defines the difference between the original samples and estimated samples of S1 and S2. There are two types of errors: sample-wise error and the mean square error of the whole signal. The sample-wise difference is between an individual original signal sample and an estimated sound signal. The optimization algorithm used is the gradient descent algorithm. Here, the target is to minimize the mean square error at each mic of the whole signal. Mean Square Error [MSE] is calculated as shown in Eq. (12). $x^1_{orgmix}(i)$ is the ith sample of the original mixed-signal at mic 1 and $x^1_{estmix}(i)$ is the ith sample of the estimated mixed-signal at mic 1.

$$MSE^1_{mix} = \frac{1}{N}\sum_{n=1}^{N}\left[x^1_{orgmix}(i) - x^1_{estmix}(i)\right]^2 \qquad (12)$$

The step-by-step procedure of the Degree separator algorithm is shown below.

Step 1: Conversion of signal from Convolutional to Linear domain.

The process begins with converting the convolutive mixed signals into a system of linear equations, as described in Eqs. (10) and (11). This transformation enables the representation of the mixed signals as linear combinations of source signals.

Step 2: Initialization of coefficients:

Initialize these linear equations' coefficients (parameters) to small random values.

Set a learning rate (alpha), which determines the step size at each iteration. Choosing an appropriate learning rate is crucial to prevent overshooting or slow convergence.

Step 3: Cost Function Computation

In this step, the algorithm initializes the cost function. Calculate the cost function that measures the error between the predicted values and the actual target values. Here is the mean square error (MSE) as per Eq. (12). This step measures how well the coefficients match the real mixed signals.

Step 4: Cost Derivative Calculation

Calculate the gradient of the cost function concerning each coefficient. The cost derivative is computed. The coefficient values are moved by slope and direction (sign) to acquire a reduced cost on the next iteration.

Step 5: Coefficient Update

The update of the coefficient follows the rule: new coefficient = coefficient - (alpha * delta), where "alpha" is the learning rate parameter and "delta" represents the change in coefficients. A learning rate parameter controls the modification of the coefficients in each iteration until the cost of the coefficients is close to the threshold set by the user.

Step 6: Back Substitution in the synthesis stage

After obtaining the nth sample estimates of the source signals S1 and S2, the algorithm reintroduces these values into the synthesis stage, creating a new set of linear equations for each microphone. The algorithm iterates through these steps until the cost function is minimized.

This iterative optimization process enables the Degree Separator to learn and adjust its coefficients to achieve the best possible estimates of the source signal samples and effectively separate them from the mixed signals.

Various techniques can be employed to evaluate the performance of CNN-DOA and sound source separation with a degree separator. For CNN-DOA, metrics like Mean Absolute Error (MAE), DOA accuracy, and confusion matrix. In sound source separation, evaluation involves Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Artifact Ratio (SAR) and perceptual metrics. These metrics assess separated sources' quality and perceptual quality. We have disused evaluation more extensively in section 4.

4. RESULTS AND DISCUSSIONS

Performance evaluation of the proposed method is undertaken in both stages, first with CNN-DOA and second with the Degree separator. All the presented results are based on averaging the outcomes over 125 random test samples of 20 ms time frames in each class. For example, in the two-source case, the total number of audio mixing frames examined for evaluation is N = 125 × 21, where 21 relates to the number of possible classes and 125 relates to number of test samples per class. A test samples are extracted from the test audio mixture based on multiple source types like different male speakers, female speakers, musical audio signals, monotone, and WGN signals for stationary and moving source scenarios in various acoustic conditions.

4.1 Performance evaluation of CNN-DOA method

In our method, the number of active sources is required before the CNN –DOA training. Here, the number of sources in a given mixed signal is assumed based on ground truth information and based on this CNN –DOA training the data set is labeled. A Uniform linear array (ULA) with a DOA range of 0°-180° and a 30° resolution is used for all experimental evaluations First, we evaluated the CNN-DOA performance with different experiments using an audio recording of simulated RIR data and actual room RIR environment data. A White Gaussian Noise (WGN) signal was created using audio city software, and a Libri speech clean database was used for evaluation. For testing, randomly selected audio signals of different male speakers, female speakers, musical audio signals, monotone, and WGN signals that were created were used. Different possible combinations of sound sources with different angular positions were used to create the audio mixture and introduce signal variation during training and testing. The BSS system is blind to source type. i.e., DOA estimation is independent of source signal type. As mixing is considered convolutive mixing, recoding can have a stable sound effect during the middle portion of the audio mixture. Test recording was selected for evaluation by removing 0.4 s at the front and the end portions of the audio mixture. Final DOA estimation was done, averaging DOA results of each frame for performance evaluation parameters. The performance of the proposed CNN -DOA method is examined with Multiple Signal Classification (MUSIC) [23]. In MUSIC, Each STFT frame's pseudo-spectrum is computed at each frequency sub-band, with a 30° angular resolution over the whole DOA space. Averaging all of the time frames of a test signal gives the final DOA test signal.

The Mean Absolute Error MAE (°) of each time frame is given by:

$$MAE_{TF}(^\circ) = \sum_{m=0}^{M} |\theta_m - \theta'_m| \qquad (13)$$

where, *M* is the number of the active sound sources (i.e., case

M=2 or 3). The true and estimated DOAs of the m[th] source are denoted by θ_m and θ'_m respectively for a given time frame. Indexing to each source starts with the lowest to higher DOA values like the source S_1 with DOA θ_1 and source S_2 with DOA θ_2 and so on. The assumption is that the estimated lower DOA belongs to the first source and second lowest belongs to the second source, and so on. The MAE (°) of the given test signal is computed by averaging the MAE of each time frame in the given test signal.

Considering N is the total number of time frames of the given test speech mixture under evaluation, the accuracy of the estimated DOA in percentage (DOA-Acc.) is given by:

$$\text{DOA} - \text{Acc.}(\%) = \frac{N'}{N} \times 100 \quad (14)$$

where, N' denotes the number of time frames with accurate DOA in a given test speech mixture.

We evaluate the performance of the proposed CNN-DOA model for different types of sounds, both known and unknown. Models are trained with simulated RIRs and real RIRs for stationary and moving sources.

4.1.1 Testing of CNN –DOA model trained with simulated RIR

To evaluate the performance of the DOA- CNN method trained with simulated RIR, seen and unseen sound sources like speech signals from LIBRI and WGN are created using Audacity with different SNR levels. We consider the different acoustic condition variations, room dimensions, Reverberation Time, ULA, and microphone positions, as shown in Table 1. We assume each source is a point source signal, neglecting noise created by diffuse sources in the room and outside the room. We have three different cases, namely, one source, two sources, and three sources. One source case constitutes 7 different cases based on seven locations with 30° angular separation from 0° -180°. For the two-source case, the total number of audio mixing frames examined for evaluation is N = 125 × 21, where 21 relates to the number of possible angle combinations with 30° angular separation between the two speakers in a range of 0° -180°. Similarly, For the three source case, the total number of mixing frames examined for evaluation is N = 125 × 35, where 35 is the number of possible angle combinations with 30° angular separation between the three speakers in a range of 0° -180°.

Figure 4 shows one of the sample CNN -DOA model Confusion Matrix, which shows the performance of CNN - DOA in single-source localization. The overall accuracy of this model is 98 % in the single-source environment, which is very high compared to other models of localization like MUSIC.

The performances of the two source and three source models are evaluated for different input SNRs, namely, 5 dB, 15 dB, and 25 dB levels. The results are based on the average of all possible location combinations and different types of sound sources with different SNR levels. The performance evaluation parameters for two different cases are presented in, Table 3 and Figure 5 from the results, the proposed CNN–DOA method can provide accurate localization performance. The model works for two source and three source cases with DOA estimation accuracy of 79%, 92%, and 98% for the two source model and 75%, 89%, and 95% for the 3 source model with input SNRs 5, 15, and 25 dB, respectively. The proposed CNN-DOA technique has a significantly greater DOA accuracy and a much lower MAE than the MUSIC technique.

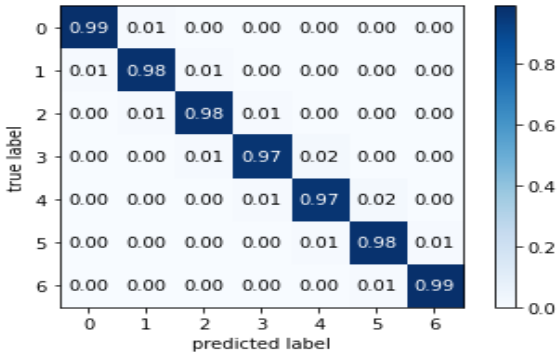

Figure 4. Confusion matrix of one of the samples CNN–DOA (Seven DOA classes in single-source 30° angular separation from 0°-180)

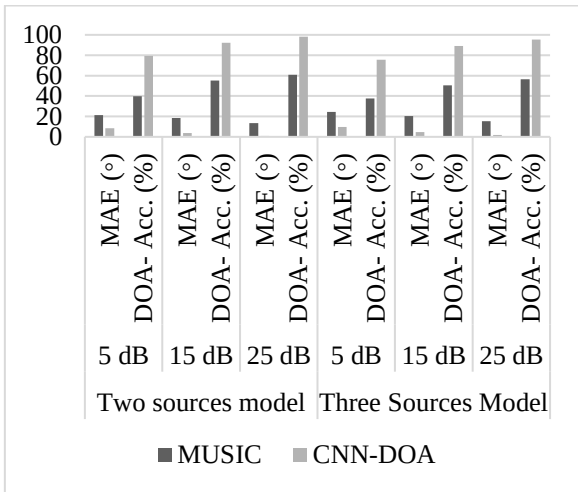

Figure 5. Performance evaluation CNN–DOA model trained with simulated RIR with two and three sources

Table 3. Performance evaluation CNN–DOA model trained with simulated RIR with two and three sources

Test Case	**Two Sources Model**					
SNR	5 dB		15 dB		25 dB	
Parameters	MAE (°)	DOA- Acc. (%)	MAE (°)	DOA- Acc. (%)	MAE (°)	DOA- Acc. (%)
MUSIC	21.2	39.8	18.4	55.1	13.5	60.8
CNN-DOA	8.5	79.5	3.8	92.1	0.8	98.2
Test Case	**Three Sources Model**					
SNR	5 dB		15 dB		25 dB	
Parameters	MAE (°)	DOA- Acc. (%)	MAE (°)	DOA- Acc. (%)	MAE (°)	DOA- Acc. (%)
MUSIC	24.3	37.6	20.3	50.6	15.2	56.6
CNN-DOA	9.8	75.5	4.7	89.2	1.9	95.5

4.1.2 Testing of CNN -DOA Model trained with real RIR for stationary and moving sources

The Multichannel Impulse Response Database from Bar-Ilan University was used in the case of measured RIRs. Experiments with CNN -DOA model trained with real RIR using Multichannel RIR Database were performed in the acoustics lab of Bar-Ilan University [19]. To evaluate the performance of DOA- CNN method trained with simulated RIR seen and unseen sound sources like speech signals from LIBRI and WGN were created using Audacity with different SNR levels. The database consists of RIRs measured in the room, as shown in Table 2 with seven different spatial source positions with an angular step size of 30 ° from 0° to 180° for a single source, two source, and three source environments. We assume each source is a point's source signal neglecting noise created by diffuse sources in the room. The total number of audio mixing frames examined for evaluation is N = 125 × 21 where 21 relates to the number of possible angle combinations with 30° angular separation between the two speakers in a range of 0° -180°. Similarly, For the three-source case, the total number of mixing frames examined for evaluation is N = 125 × 35, where 35 is the number of possible angle combinations with 30° angular separation between the three speakers in a range of 0° -180°.

Table 4. Performance evaluation CNN–DOA model trained with real RIR with two and three sources with stationary and moving sources

Test Cases	Two Source Model		Three Source Model	
Objective Measure	MAE (°)	DOA- Acc. (%)	MAE (°)	DOA- Acc. (%)
MUSIC (With stationary sources)	13.4	58.4	15.3	55.6
CNN-DOA (With stationary sources)	3.4	90.2	3.9	87.5
CNN-DOA (With moving sources)	6.4	82.1	9.1	75.6

Results for two cases with the two-source model and the three-source model of CNN - DOA and the results shown here are based on an average of all possible location combinations and different sound sources for each input SNR. The performance of three cases is evaluated and presented in Table 4 and Figure 6 based on the average of all possible location combinations and different types of sound sources for each input at different SNR levels. We evaluate the performance of the CNN-DOA model using real-world stationary and moving sound source signals with different SNR levels. The number of mixtures frames that are under evaluation during testing is N = 125 ∗ 10 = 1250, where 10 corresponds to a randomly selected combination in two sources of three sources with 30° angular separation between the sources in a range of 0°-180°, e.g., in the two-source case the S1 stationary source is at S1-30° and the S2 source is at 60° moving toward 90°. In the three-source case, the S1 stationary source is at S1-30°, the S2 source and the S3 source are at 90° moving toward 120°. As shown in Table 4 and Figure 6, the proposed CNN – DOA method can provide accurate localization performance compared to MUSIC in stationary and moving sources in actual room environments. The proposed model shows a DOA estimation accuracy of 90.2% in the two-source model and 87.5% for the three-source model in the stationary case and a little lower in the moving source case. The overall accuracy of CNN DOA decreases with stationary to moving source case as the model is trained only in stationary source cases.

In actual room setup, CNN-DOA has higher overall accuracy than MUSIC in stationary and moving cases. Here, stationary and moving source conditions are used to create a CNN-DOA training database, strengthening the model under both conditions. The tricky part of training CNN-DOA with moving sources is creating a reliable model that reduces reverberation and Doppler shift. However, CNN-DOA-based sound source separation is a promising approach for moving sources compared to the MUSIC technique. The proposed method's accuracy in moving sources can be improved by adding temporal information and adaptive source tracking.

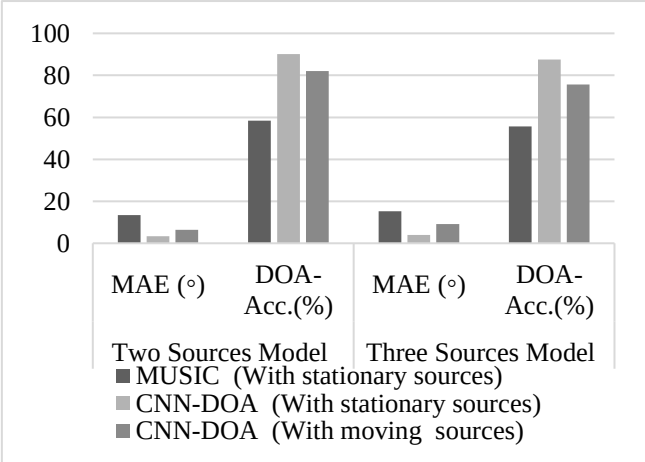

Figure 6. Performance evaluation CNN –DOA model trained with real RIR with two and three sources with stationary and moving sources

4.2 Performance evaluation of degree separator

In this section, the proposed degree separator separation performance has been evaluated. We have compared the performance of the proposed degree separator with conventional BSS methods like FICA. The performance of degree separator methods is evaluated in simulated RIR and recoded RIR in a room with speech signals from LIBRI and other sound sources with different SNR levels. A separation performance evaluation is based on objective measures such as the image-to-spatial distortion ratio (ISR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) [16].

The test mixtures were generated by the convolution of the source-to-array RIRs with different anechoic source signals with male and female speech samples, music, and single-frequency tones as different sources. Array mixtures are created based on adding spatial images of each source at a specific location for two and three source conditions. In the degree separator, we need to set the following parameters: window length, learning rate, and the number of iterations.

The BSS Evaluation toolbox [16] separated the signal with SDR, ISR, SAR, and SIR as four objective parameters. The SDR calculates how much of the original signal has been retrieved by the estimated signal. The SIR determines the amount of interference created by other sources in the targeted signal. It measures the performance of separation of the target source in multiple source environments assuming other sources as interference. The SAR measures additional artifacts produced by the separation process, and the ISR measures how

the algorithm preserves the spatial image of the estimated source signal after reconstruction. The score of each time frame of the test signal segment in the dB scale is converted into a linear scale and averaged over the all-test segments and again converted into the dB scale. The resulting matrixes are shown in Table 5 and Figure 7, consisting of an average of these all-different cases and adding different possible combinations of the source locations in the room environment with two and three source experimental conditions.

Table 5. Performance evaluation of degree separator in two and three source model with simulated RIR and recorded RIR

Test Cases	Model	Two Sources Model			
Experimental Condition	Objective Measure	ISR (dB)	SIR (dB)	SAR (dB)	SDR (dB)
Simulated RIR and Recorded RIR	FICA	6.9	4.4	6.2	2.3
	Degree Separator	9.2	8.3	11.1	6.7
Test Cases	**Model**	**Three Sources Model**			
Experimental Condition	Objective Measure	ISR (dB)	SIR (dB)	SAR (dB)	SDR (dB)
Simulated RIR and Recorded RIR	FICA	4.9	2.5	5.7	1.6
	Degree Separator	8.1	7.5	10.3	4.67

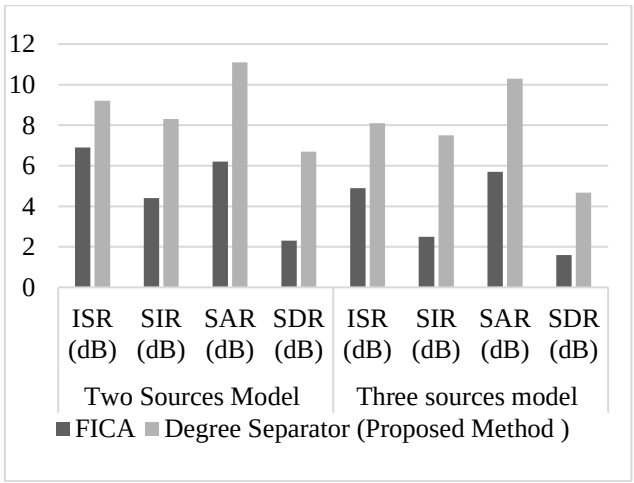

Figure 7. Performance evaluation of degree separator in two and three source models with simulated RIR and recorded RIR

The results demonstrated in Table 5, and Figure 7 show that SDRs and SIRs are higher with the degree separator than in traditional FICA approaches. In all test sets, the proposed method is the most effective at reconstructing the spatial image of the source. The proposed method provides two simultaneous source separations better than three simultaneous sources. The SAR score shows that added artifacts to the separated signals are lower than the FICA score in comparison to the proposed method. SDR values for both two sources and three sources are higher for the degree separator than the FICA. The proposed method exceeds ICA-based separation in different source mixing cases such as male speech, female speech, music, and single tone as the source.
Separation quality based on perception was measured using Short-Term Objective Intelligibility (STOI) [23]. The higher the STOI, the better the speech separation and a superior value of STOI is around 0.9. We compared the performance of source separation using FICA and Degree separator in terms of STOI in separated signals from mixing different sources such as male speech, female speech, music, single tone etc., as shown below in Table 6 and Figure 8.

Table 6. Performance evaluation of degree separator using STOI (Avg.) in the two and three sources model

Model	Two Sources Model		Three Sources Model	
Experimental condition	Simulated RIR	Recoded RIR	Simulated RIR	Recoded RIR
FICA	0.64	0.61	0.58	0.55
Degree Separator	0.86	0.83	0.81	0.78

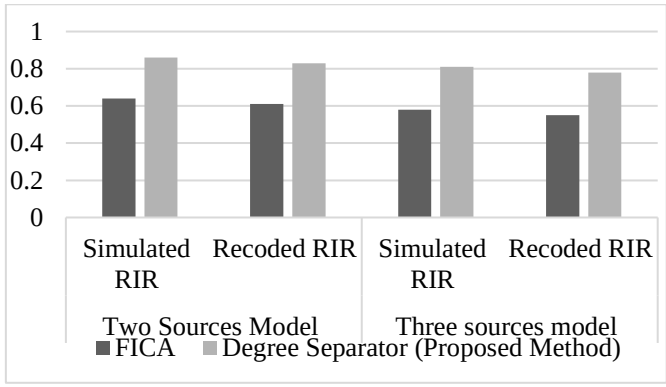

Figure 8. Performance evaluation of degree separator using STOI in the two and three sources model

5. CONCLUSION

The novel method presented in this research significantly advances the field of sound source separation by introducing a combined approach that employs CNN-based DOA estimation in conjunction with the innovative degree separator. The database for training and testing of CNN-DOA and degree separator was created with up to three sound sources, including two moving sources. This model is trained using simulated and actual room-recorded databases for both cases, i.e., moving and stationary sources. This research proposes a novel technique combining CNN-DOA and a new degree separator technique to separate sound sources. The performance of the degree separator is evaluated with the help of BSS evaluation parameters. The result shows that the proposed approach improves SDR, SIR, SAR, and ISR compared to previously available approaches like FICA. The research demonstrates that the proposed method of CNN-DOA (with the Degree source separator) shows high practical applicability for BSS separation. It confirms the effectiveness of the DOA estimation and separation quality of signals for simulated and actual room recordings compared with FICA. Improved source separation in a room environment has significant real-world applications, such as more accurate speech recognition systems, immersive sound experiences in music production, enhancing forensic analysis. However, we acknowledge certain limitations, including the need to accurately determine the number of active sources, separation performance and computational cost in more complex auditory scenes, variation in room geometry, sensitivity to external noise, and separation performance with more moving sources. Researchers can also explore source separation performance through a dynamic source tracking mechanism, training CNN-DOA with multiple real-room environments, and adding a

noise filter at the preprocessing stage of separation. These prospective research endeavors have the potential to refine this method further, making it even more applicable in real-world and potentially revolutionizing the field of audio source separation technology.

ACKNOWLEDGMENT

This work is supported by the Centre of Excellence in Signal and Image Processing-(CoE-S&IP) at the College of Engineering, Pune. CoE-S&IP is funded by MHRD-World bank under TEQIP-II.

REFERENCES

[1] Vincent, E., Virtanen, T., Gannot, S. (Eds.). (2018). Audio Source Separation and Speech Enhancement. Wiley Publication.

[2] Nikunen, J., Virtanen, T. (2014). Direction of arrival based spatial covariance model for blind sound source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(3): 727-739. https://doi.org/10.1109/TASLP.2014.2303576

[3] Hyvarine, A., Karhunen, J., Oja, E. (2001). Independent Component Analysis. John Wiley & Sons Publication Inc. https://doi.org/10.1002/0471221317.fmatter_indsub

[4] Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T., Shikano, K. (2003). Blind source separation combining independent component analysis and beamforming. EURASIP Journal on Advances in Signal Processing, 2003: 1-12. https://doi.org/10.1155/S1110865703305104

[5] Mukai, R., Sawada, H., Araki, S., Makino, S. (2004). Frequency domain blind source separation for many speech signals. In: Puntonet, C.G., Prieto, A. (eds) Independent Component Analysis and Blind Signal Separation. ICA 2004. Lecture Notes in Computer Science, vol 3195. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30110-3_59

[6] Ikram, M.Z., Morgan, D.R. (2002). A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation. In 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, pp. I-881-I-884. https://doi.org/10.1109/ICASSP.2002.5743880

[7] Hoshuyama, O., Sugiyama, A., Hirano, A. (1999). A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. IEEE Transactions on Signal Processing, 47(10): 2677-2684. https://doi.org/10.1109/78.790650

[8] Meyer, J., Elko, G. (2002). A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield. In 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, pp. II-1781-II-1784. https://doi.org/10.1109/ICASSP.2002.5744968

[9] Wang, L., Reiss, J.D., Cavallaro, A. (2016). Over-determined source separation and localization using distributed microphones. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(9): 1573-1588. https://doi.org/10.1109/TASLP.2016.2573048

[10] Nikunen, J., Diment, A., Virtanen, T. (2017). Separation of moving sound sources using multichannel NMF and acoustic tracking. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(2): 281-295. https://doi.org/10.1109/TASLP.2017.2774925

[11] Ozerov, A., Févotte, C., Vincent, E. (2018). An introduction to multichannel NMF for audio source separation. In: Makino, S. (eds) Audio Source Separation. Signals and Communication Technology. Springer, Cham. https://doi.org/10.1007/978-3-319-73031-8_4

[12] Innami, S., Kasai, H. (2012). NMF-based environmental sound source separation using time-variant gain features. Computers & Mathematics with Applications, 64(5): 1333-1342. https://doi.org/10.1016/j.camwa.2012.03.077

[13] Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Transactions on Audio, Speech, and Language Processing, 15(3): 1066-1074. https://doi.org/10.1109/TASL.2006.885253

[14] Mirzaei, S., Norouzi, Y. (2015). Blind audio source counting and separation of anechoic mixtures using the multichannel complex NMF framework. Signal Processing, 115: 27-37. https://doi.org/10.1016/j.sigpro.2015.03.006

[15] Chakrabarty, S., Habets, E.A. (2019). Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. IEEE Journal of Selected Topics in Signal Processing, 13(1): 8-21. https://doi.org/10.1109/JSTSP.2019.2901664

[16] Vincent, E., Gribonval, R., Févotte, C. (2006). Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing, 14(4): 1462-1469. https://doi.org/10.1109/TSA.2005.858005

[17] Ozerov, A., Févotte, C. (2009). Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Transactions on Audio, Speech, and Language Processing, 18(3): 550-563. https://doi.org/10.1109/TASL.2009.2031510

[18] Habets, E.A.P. (2006). Room impulse response generator. Acoustic Sensor Networks-Geometry Calibration View project Artificial Reverberation View project. https://www.researchgate.net/publication/259991276.

[19] Hadad, E., Heese, F., Vary, P., Gannot, S. (2014). Multichannel audio database in various acoustic environments. In 2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC), Juan-les-Pins, France, pp. 313-317. https://doi.org/10.1109/IWAENC.2014.6954309

[20] Chakrabarty, S., Habets, E.A. (2017). Multi-speaker localization using convolutional neural network trained with noise. arXiv preprint arXiv:1712.04276. https://doi.org/10.48550/arXiv.1712.04276

[21] Juneja, S., Juneja, A., Dhiman, G., Behl, S., Kautish, S. (2021). An approach for thoracic syndrome classification with convolutional neural networks. Computational and Mathematical Methods in Medicine, 2021: 3900254. https://doi.org/10.1155/2021/3900254

[22] Dmochowski, J.P., Benesty, J., Affes, S. (2007). Broadband MUSIC: Opportunities and challenges for multiple source localization. In 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, pp. 18-21. https://doi.org/10.1109/ASPAA.2007.4392978

[23] Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing, 19(7): 2125-2136. https://doi.org/10.1109/TASL.2011.2114881

NOMENCLATURE

A	Magnitude component
b	Frequency bin in STFT of signal
h	Room impulse response
k	Number of time frames of signal
m	Number of microphones
N	Number of time frames under testing
S	Source signal
Xmix	Mix signal

Greek symbols

θ	Direction of arrival of source signal
ϕm	Phase component in STFT