IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 November 2022, accepted 28 November 2022, date of publication 1 December 2022,
date of current version 6 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3225968

RESEARCH ARTICLE

Multiple Sound Source Localization in Three Dimensions Using Convolutional Neural Networks and Clustering Based Post-Processing

SAULIUS SAKAVIČIUS, (Member, IEEE), ARTŪRAS SERACKIS, (Senior Member, IEEE), AND VYTAUTAS ABROMAVIČIUS, (Member, IEEE)
Faculty of Electronics, Department of Electronic Systems, VILNIUS TECH, 10221 Vilnius, Lithuania
Corresponding author: Vytautas Abromavičius (vytautas.abromavicius@vilniustech.lt)

ABSTRACT Sound source localization methods are successfully applied for various estimation tasks, such as tracking and detecting objects, aiming cameras, and navigating robots. However, large and usually complex distributed microphone arrays are used for three-dimensional acoustic source localization. This study proposes a convolutional neural network architecture for three-dimensional sound source localization using a single tetrahedral microphone array. A spectrum phase component of a microphone array signal was designed as the input of the model, while the output represents a three-dimensional space. The paper provides extensive experimental results of the given method on a semi-synthetic audio data set and a real-world microphone array. Furthermore, cluster-based post-processing has been shown to increase the accuracy of three-dimensional localization by more than 30%. The experimental results on a synthesized data set using the image source method showed 1.08 m localization uncertainties. The estimate of the investigated sound sources had a mean absolute error of 18.97° and elevation error of 48.49°. An additional advantage of the proposed method is the ability to predict the location of the sound source from a single signal analysis frame. This gives instant localization and is in line with many alternative applications. The proposed solution does not require intensive preprocessing of the audio signals and can be used as a video camera pointing system based on a microphone array. In the future, it would be relevant to investigate the localization performance of more than two sound sources, and the variable acoustic conditions could also be assessed.

INDEX TERMS Clustering, convolutional neural networks, microphone array, multiple sound sources, sound source localization, three-dimensional.

I. INTRODUCTION

Sound source localization (SSL) is an important topic in human-machine interface, robotics [1], [2], security [3], and autonomous driving [4]. Applications of sound source localization in three-dimensional (3D) space include event detection and tracking, microphone array beamforming or camera aiming, and robotic navigation [1], [5].

The direction of arrival (DoA) of one or more active sound sources can be used to steer the directivity pattern of a microphone array in ambient intelligence systems [6] or security surveillance systems [3]. If the acoustic localization system is in a different position than the system that uses the localization information, the DoA estimation is not sufficient, and complete 3D coordinates of a sound source are needed. For example, assume that there is a surveillance system composed of an acoustic source, a microphone array, and a camera. If the microphone array is at a different position from the camera that needs to be aimed at the estimated source position, one would need to compensate for the parallax error, and this is only possible if the exact source position is known (as opposed to the only known DoA of the source), as well as the positions of the camera and the microphone array (see Fig. 1).

Currently, there are an abundance of source DoA estimation algorithms, such as simple time difference of arrival (TDoA)-based estimation algorithms, such as Generalized

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

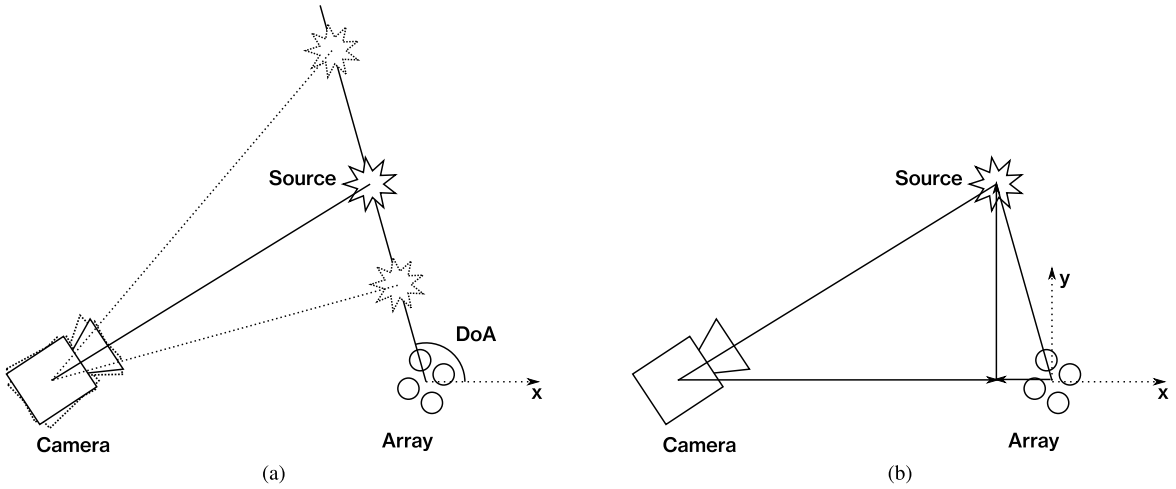


FIGURE 1. Uncertainty of camera aiming due to unknown distance between the source and the array (a); certain camera aiming when the complete set of source coordinates is known (b).

Cross-Correlation with Phase Transformation (GCC-PHAT). More complex and more robust algorithms, such as Steered Response Power Phase Transform (SRP-PHAT) [7], which are designed for adverse acoustic environments. And, complex but accurate statistical signal processing-based Minimum Variance Distortion-less Response (MVDR) beamformer or eigenvalue decomposition-based Multiple Signal Classification (MUSIC) algorithm or Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT).

The methods discussed are well suited to estimate the DoA of single or multiple sound sources. Estimating the azimuth and elevation of the sound source and the distance to the microphone array of a sound source within an acoustic scene is much more complicated. This is mainly because the TDoAs of sound sources at the same DoA but at different distances are very close. For TDoA based source locators, it is virtually impossible to discern between near and far sources without using additional features such as the Direct-to-Diffuse ratio. Estimating the source-array distance becomes easier when the array aperture is increased because the TDoAs between microphones that are spaced further apart are also greater. In this context, distributed microphone arrays or microphone networks are better suited for three-dimensional source localization. However, such arrays offer lower portability and higher complexity.

The 3D sound source localization is not a new concept. There have been several investigations regarding the localization of 3D sound sources. Most of the proposed methods employ multiple spatially distributed microphones (an arbitrarily shaped large aperture array [8], [9], [10]) or microphone arrays [11], but such systems are relatively complex and are not portable. Only a few attempts were made to discern the position of the source using a single compact microphone array. The compact microphone array is defined as an array having a much smaller aperture than the dimensions of the enclosure (or search space).

Recently, multiple investigations have been presented on the application of Artificial Neural Networks (ANN) for SSL [1], [12]. Learning-based SSL methods show advantages in situations where complex functions define the relationship between the microphone array signals and features extracted from those signals, or where the positions or DoAs of sound sources are approximated. In most cases, the Convolutional Neural Network (CNN) [13], [14] and the Recurrent Neural Network (RNN) [15] or a combination of both are used [16], [17], [18], [19], [20]. In addition, auto-encoders have also been investigated in several works [21], [22], [23]. In general, SSL using ANN is commonly formulated as a classification [14], [18], [24], [25], [26], [27], [28], [29], [30], or a regression problem [16], [31], [32], [33].

In the case of the regression problem, the ANN output is a vector of one, two, or three dimensions (for a single sound source localization [22], [34]) or a set of vectors (for the localization of multiple sound sources [35], [36], [37]). In the case of the classification problem, the input features are classified into an array of spatial classes that represent the source coordinates in one, two or three dimensions.

In the case of one-dimensional sound source localization, the estimated dimension is usually the azimuth (DoA [19], [21], [25], [38], [39]). In the case of two-dimensional (2D) sound source localization, the estimated dimensions are usually azimuth and elevation (2D DoA [40], [41], [42]). In the case of SSL in three dimensions, the proposed methods estimate the polar [18] or Cartesian [43], [44], [45], [46] coordinates. It should be noted that sound source localization in three dimensions is generally approached as a regression problem, and only in a few cases has it been investigated as a classification problem [18], [47].

Generally, for the supervised training of an ANN, a large data set of labeled samples is needed, which is costly to acquire. Additionally, in the case of SSL, the data set needs to contain a number of multichannel array signal recordings

with the positions of the sound source labeled for each frame of the recording. Few data sets are publicly available [48], [49]. Another way to obtain an SSL data set is to simulate the array signals using acoustic models. Most often, an image source method [50] is used to estimate the impulse response of a room, and then multipath propagation and associated effects can be introduced. Although SSL in a reflection-free environment is often an easier task, most authors investigated SSL in reverberating environments because such a scenario is more common in real-life situations.

Various input features were proposed to be used with ANN-based sound source localization methods, such as interaural level, phase or time difference [16], [18], [51], [52], phase transform-based features [38], [53], [54], [55], [56], magnitude and phase spectrograms of array signals [35], [41], [57], [58], [59] and even unprocessed audio waveforms [20], [60], [61], [62], [63], [64], [65], [66].

Chakrabarty et al. [25] proposed a method for the estimation of DoA of multiple acoustic sources (azimuth only) using a CNN. Phase components of the Short-time Fourier transform (STFT) frame of the source signal were used as CNN input features, and a vector representing the posterior probability of a sound source was used as the desired output during CNN training. Furthermore, Laufer-Goldshtein et al. [67] in their work showed that multidimensional acoustic features lie in a manifold embedded in a low-dimensional space and that these features exhibit spatial smoothness [67]. The investigation showed that sound sources that are spatially close have acoustic features that are also close in the embedded low-dimensional space.

For the ANN, spatial smoothness is well learned using a target format, where the probabilities of the adjacent classes also exhibit smoothness. For example, by using Gaussian blurring of the target, ANN learns the mapping much more efficiently and accurately. CNN in the works of He et al. [38] and Chakrabarty et al. [25] is used as a classifier that classifies input features into classes that represent the DoA of the sound source(s). To improve the learning of the classification of spatially smooth acoustic features, He et al. [38] used Gaussian smoothing of the CNN output feature, while a similar smoothing of the output vector can be observed in Chakrabarty et al. [25] work, where CNN was implicitly trained.

In the context of the methods described, our proposed method is intended to estimate the three-dimensional coordinates of multiple wide-band (speech) sound sources that are stationary for the duration of the analysis time frame. We formulate the source position estimation task as a position classification task. We propose to use the STFT phase component of the microphone array signal frames. The main contribution of this paper is the proposed method, which uses a three-dimensional matrix as the CNN output. The values of the matrix elements represent the posterior probability of a sound source being active at a particular spatial position. The elements of the proposed output matrix are interpreted as spatial classes to which the CNN classified the STFT input features. A spatial class represents a particular set of coordinates in a three-dimensional space. We extend our previous research on 2D DoA heatmap estimation using CNN and STFT phase input features [68], [69] and advance the thesis idea on multiple sound source localization methods [70]. Essentially, we propose a clustering-based method for the acquisition of the sound source coordinates in vector form from the three-dimensional output matrix.

II. MATERIALS AND METHODS

In this section, we present the methods to acquire the STFT phase input feature, the desired 3D output matrix, and the CNN structure used for the investigation. We present the maximum matrix element and a clustering-based source coordinate estimation procedure. We also provide a description of the acquisition of training and testing data sets, the CNN training procedure, and the evaluation of the performance of the proposed method.

A. CONSTRUCTION OF INPUT FEATURES

The preparation of the input features is carried out in several steps. First, the acoustic signals within an enclosure are captured using a non-coplanar microphone array $\mathbf{M}$ consisting of N_M microphones. The signals are then converted to a digital representation of the signal using analog-to-digital converters (ADCs) at a sampling rate f_s. In this investigation, $N_M = 4$ and $f_s = 16$ kHz.

In the next step, the STFTs of the microphone signals are calculated. First, the microphone signals are framed to 0.1 s duration frames. For each frame and for each of the $N_M = 4$ microphone channels, the Fast Fourier Transform (FFT) is calculated. The number of FFT points equal to $N_{\text{STFT}} = 512$, with 256-point overlap and a Hann windowing function. The number of frequency bins in the STFT was $N_f = N_{\text{STFT}}/2 + 1 = 257$. As a result, an array of size $N_M \times N_f$ is obtained. As input feature, a single STFT frame is used, a matrix with $N_M \times N_f = 4 \times 257$ elements.

Examples of the prepared input features are presented in Fig. 5. In the figure, STFT magnitude and phase feature examples of noise and speech signals, 1 and 2 simultaneously active acoustic sources; 4 channel (tetrahedral) microphone array; input features are the STFT phase component.

Extending the work of Chakrabarty et al. [25] we used the phase component of the STFT calculated for microphone array signals. However, we did not explicitly take into account the W-disjoint orthogonality of the signals. According to Chakrabarty et al. [25], in the case of a N_S-source scenario, for each of the sources, the array signals are simulated using the image-source method separately. Then the STFTs of the receiver signals are concatenated and randomly permuted in both time and frequency domains (leaving only the channel order unchanged). We, on the other hand, do not perform the permutation of the frequency or channel dimensions of the STFT feature and obtain microphone array signals and, in turn, the STFT phase input features with all sources present and active within the enclosure at once.

B. DESIRED OUTPUTS

The elements of a 3D matrix of the proposed method are used as a desired output for each input feature matrix. The matrix is an array of $K \times L \times M$ elements, where each element represents a point in the physical space, and the value of the element represents the posterior probability that a sound source is active at that point in the space. The volume covered by the 3D matrix elements is chosen arbitrarily and in our investigation coincided with the volume of a cuboid-shaped acoustic enclosure. The number of elements in the 3D matrix along the x, y and z axes, respectively, can be expressed in terms of the density of elements per unit of length Q_x, Q_y and Q_z, which represent the spatial resolution of the 3D matrix, and the lengths of the sides of the volume, respectively X, Y and Z (in meters), which is represented by the 3D matrix:

$$[K, L, M] = [X, Y, Z] \circ [Q_x, Q_y, Q_z], \tag{1}$$

here ($\circ$) denotes the Hadamard product of the vectors. In our investigation, the spatial resolution of the output matrix was equal on all axes: $Q_x = Q_y = Q_z = Q$.

A target feature for CNN training was generated in the following steps:

1) An empty 3D matrix was created. The number of elements in the matrix along each axis defines the spatial resolution of the target feature.
2) A 3D Gaussian kernel function was evaluated using the 3D matrix with the center of the kernel positioned at $\mathbf{s} = [s_x, s_y, s_z]$. The spread of the Gaussian kernel σ determines the spatial smoothness factor of the target feature. A 3D Gaussian kernel has 3 spread values, one along each axis: σ_x, σ_y, σ_z. In our investigation, spread values along all axes were the same: $\sigma_x = \sigma_y = \sigma_z = \sigma$.
3) The steps are repeated N_S times.

The resulting 3D matrix contains a Gaussian kernel with a particular σ placed at the coordinates of each simulated sound source. Afterwards, a CNN was trained to estimate such 3D matrices from the provided STFT phase input features. Acoustic features exhibit spatial smoothness that is reflected in the feature space [67]. In contrast, an ANN is expected to classify such input features as neighboring classes in the output. Therefore, we speculate that the 3D matrix blurring operation would allow CNN to learn to map features that are nearby in the feature space to neighboring spatial classes. The values of the output layer of the ANN represent the posterior probability of a feature being obtained for a sound source at a particular point in space. A feature for a source at a particular spatial position can be viewed as having a lower but non-zero posterior probability of being obtained for a source at a slightly different (neighboring) position. Thus, we believe that this angular smoothing of the 3D matrix would be beneficial for learning the ANN as well as its robustness against multipath propagation.

Examples of the desired output prepared at, respectively, $Q = 0.5$ m and $\sigma = 1$ m with a single active source and $Q = 0.25$ m and $\sigma = 0.5$ m with two active sources are presented in Fig. 2. In the figure, a 3D grid of points is shown, where each point represents a spatial class that is associated with a particular location within a three-dimensional space of the acoustic enclosure. Each point was assigned a value in the range of [0; 1], to represent the probability that the sound source being active at a particular location. This three-dimensional representation of a volume within an acoustic enclosure was used as a training target (desired output) for the CNN, which was expected to learn the mapping between the STFT phase input feature and the aforementioned training target.

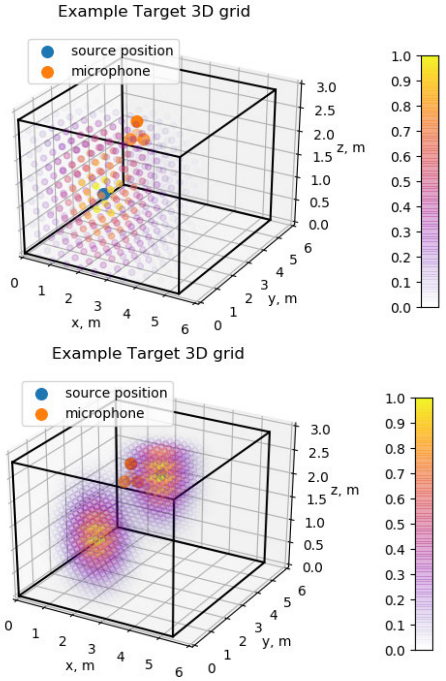


FIGURE 2. Examples of the desired output with various resolution and Gaussian kernel spread values: $Q = 0.5$ m, $\sigma = 1$ m, single source with coordinates $\mathbf{s} = [1.2, 2.1, 1.3]$ m and $Q = 0.25$ m, $\sigma = 0.5$ m, two sources with coordinates $\mathbf{s_1} = [1.2, 2.1, 1.3]$ m and $\mathbf{s_2} = [3.1, 3.2, 2.3]$ m; ground truth source positions are marked with blue circles.

C. CNN ARCHITECTURE

The proposed architecture of the CNN algorithm was based on our previous research [69], which, in turn, was derived from [25].

Our newly proposed CNN architecture consists of three 2D convolutional layers with 128, 64, and 32 units, respectively, with a convolution kernel size of (2×1) elements. The convolutional layers are followed by a dropout layer with a fixed dropout rate of 0.125. The dropout layer is followed by three fully connected layers, each containing (257×4) units, again followed by a dropout layer with a dropout rate of 0.125. Finally, there are a 1028-element fully connected layer and a $K \times L \times M$ element fully connected layer that is a vector reshaped into a $K \times L \times M$ 3D matrix of elements at the CNN output, where each element of this last layer represents

a spatial position in a 3D output matrix. The reshaping of the vector into a 3D matrix is only a representation-related operation, which allows to pass a 3D feature as a training target for the neural network. The output layer of the neural network is a fully-connected layer. Thus, it does not inherently convey a 3D structure and does not explicitly carry any geometric information. The output of this layer is interpreted as a 3D structure, which is done to allow the usage of 3D matrices for training the network. The CNN is expected to learn to produce output similar to the desired output provided during the training of the network, regardless of the way such outputs are interpreted later. In summary, the output layer itself does not contain geometric spatial information, but the output is interpreted in a spatial context. It is considered here that the spatial relation between the output elements of the output layer is implicitly learned by the neural network, without having to rely on convolutional output layers. Exponential Linear Unit (ELU) as the activation function was used in all layers of the CNN. We have used binary cross-entropy as the loss function and Adaptive Moment Estimation (Adam) optimizer. The diagram of the CNN architecture is presented in Fig. 3, where the layers of the neural network are depicted as rectangles, with the number of elements along each dimension of the particular layer marked along the respective edges of the rectangles.

The number of neurons in the CNN output layers depends on the number of elements in the 3D output matrix, which, in turn, depends on the resolution Q of the spatial 3D matrix and the dimensions of the acoustic enclosure. CNN must be trained for each different Q, σ of the 3D matrix, and $[X, Y, Z]$ of the enclosure.

According to [25], the CNN architecture used with the STFT phase features of the N_M channel can have, at most, $N_M - 1$ convolution layers, since after the $N_M - 1$ layers, performing 2D convolution is no longer possible as the feature maps become vectors. Thus, in our architecture, at most 3 convolutional layers can be used. Regarding the number of convolutional layers, CNN performance was shown to improve with an increase in the number of convolutional layers to M-1. The number of fully connected layers was selected on the basis of previous research by the authors.

D. SOURCE COORDINATE ESTIMATION FROM OUTPUT 3D MATRIX

For single source localization, the coordinates of the element with maximum value were found and converted to Cartesian coordinates by dividing by the resolution of the 3D matrix.

For multiple sound source localization, a threshold was applied to the 3D matrix, removing elements that had a value lower than the mean of the entire field multiplied by a coefficient J. The remaining elements were then clustered using K-means clustering to N_S clusters. The K-means clustering algorithm assigns data points to a specified number of uniformly spread clusters based on the squared error distance. N_S can be an arbitrary number, and the algorithm is supposed to find the most probable source positions N_S. The center coordinates of the clusters correspond to the source coordinates. Thus, for the investigation N_S was set to 1 and 2. It should be noted that the clustering-based source position estimation method allows the localization of an arbitrary number of sources simply by selecting the number of clusters into which the thresholded 3D matrix elements are clustered by the K-means clustering algorithm.

E. DATA SETS

To evaluate CNN performance with various Q and σ parameters of the CNN output feature as well as the number and type of sound sources, multiple data sets were generated.

We have evaluated the performance of CNN trained on data sets with one or two simultaneously active sound sources and with $Q \in [0.25, 0.5, 1]$ m and $\sigma \in [0.25, 0.5, 1]$ m. The source positions were randomly selected within the limits of a simulated acoustic enclosure with dimensions of 5.4 m, 5.86 m and 2.84 m in x, y and z, respectively.

A tetrahedral microphone array was used, with microphone positions $\mathbf{m}_i = [m_{ix}, m_{iy}, m_{iz}]$, $i \in [1, 2, 3, 4]$ presented in Table 1.

TABLE 1. Positions of the microphones of the tetrahedral microphone array.

i	m_{ix}, [m]	m_{iy}, [m]	m_{iz}, [m]
1	3.0	2.0	2.0
2	3.4	2.0	2.0
3	3.2	2.35	2.0
4	3.2	2.12	2.35

For a particular N_S, the source positions were generated once and used to generate variants of the data set with different signals (noise or speech), Q and σ; 2 data sets in total. These are the ground-truth data sets for source positions (see Fig. 4).

For the two-source data set, the source positions were generated in sets of two source positions per set, and the audio signals of microphone the array were simulated for each of the sets of source positions.

Synthetic microphone array audio data sets with noise and speech signals were generated using the image source method, implemented in *pyroomacoustics* Python package [71]. Room impulse responses were created for particular sets of source and microphone positions using image-source method considering up to 7^{th} order reflections, and with the absorption coefficient of the enclosure boundaries set to 0.8. Anechoic speech and noise signals were convolved with the room impulse responses to obtain the simulated microphone array signals.

For the creation of the speech evaluation data set, speech signals were randomly selected for each source position in the AMI corpus [72], from a subset of dry microphone recordings of 5 seconds or more (longer records were truncated to 5 s). Noise signals were generated dynamically during data set creation; samples of these signals were randomly sampled

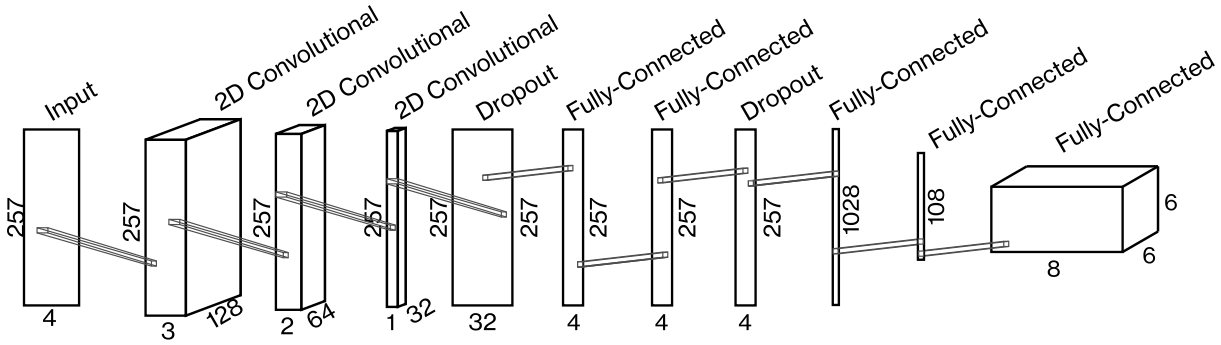
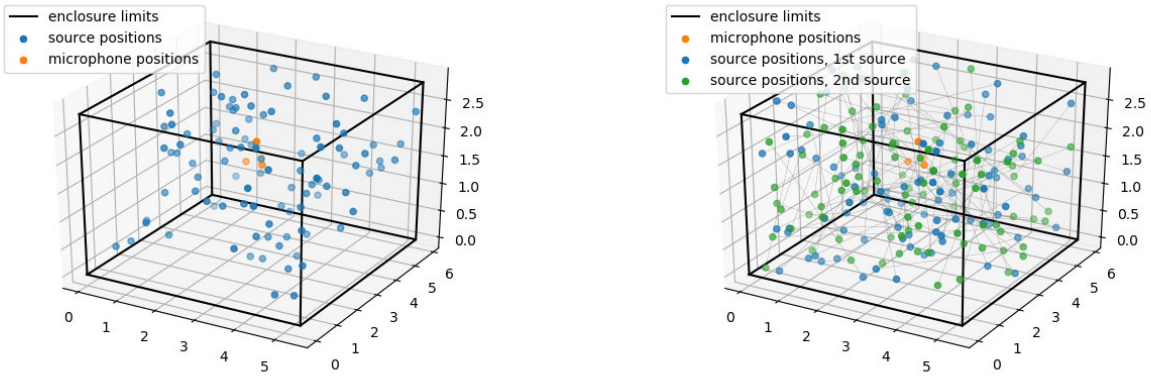


FIGURE 3. A diagram of the architecture of the CNN used for the investigations.



(a) (b)

FIGURE 4. Source positions used for single (a) and two source (b; lines connect source positions for the same position set) data set generation.

from a uniform distribution, and a gain of 0.9 was applied, creating white noise.

STFT data sets were created once for each N_S with noise and speech signals, 4 data sets in total, with parameters described in Section II-A. These are the input feature data sets.

The desired 3D matrices of CNN output features were generated for each N_S and for each Q and σ; 18 data sets in total (9 for a single source, 9 for two sources). These are the training target data sets.

Training input and target feature data sets were generated only using noise signals at 1×10^2 source positions, with one STFT frame per position, resulting in 1×10^2 training samples.

Evaluation data sets were generated using both noise and speech signals at 100 source positions with 314 STFT frames at each position, resulting in 31 400 evaluation samples. Multiple frames per single source position were generated because the speech signal is non-stationary and the prediction result for an input frame is dependent on the audio content of a particular audio frame from which the input feature was generated; thus, it is desired to evaluate each source position using more than one speech signal frame.

Furthermore, the performance of the proposed method was evaluated using an openly accessible data set of tetrahedral microphone signals [73]. This data set was used because it contains information about not only the source and microphone positions relative to the enclosure walls, but also the dimensions and acoustic properties of the enclosure, allowing one to synthesize a large training data set using the image source method to train CNN.

F. CNN TRAINING AND EVALUATION

The same CNN architecture was trained on noise signal data sets containing 1×10^5 samples that were created for each of the $Q \in [0.25, 0.5, 1]$ m and $\sigma \in [0.25, 0.5, 1]$ m as described in Section II-E. CNN was trained for at most 100 epochs due to resource limitation, with validation loss early stopping criteria. Training was carried out using a batch size of 512 samples, with the learning rate of the optimizer set to 0.001.

To evaluate the performance of each trained CNN model, testing data sets were used. CNN predicted a 3D matrix output feature for each STFT phase input feature of the data set for each STFT phase.

An additional training data set was created to evaluate the performance of the proposed method using real-world microphone array signals, simulating the geometry, acoustic properties of the enclosure, and the microphone arrays described in [73].

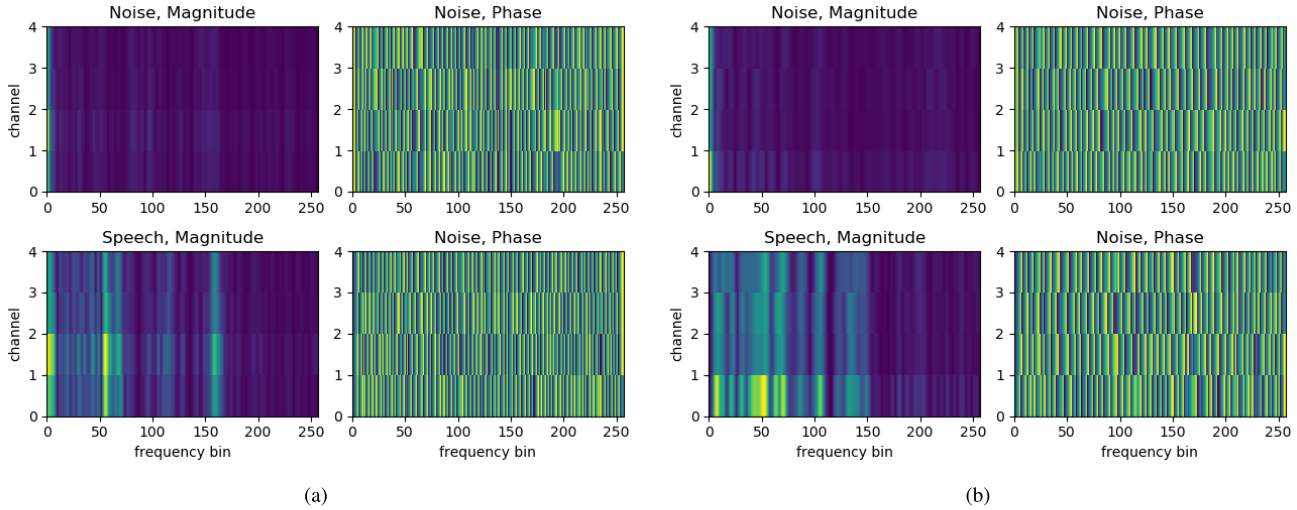


FIGURE 5. STFT magnitude and phase feature examples of noise and speech signals, 1 (a) and 2 (b) sources; 4 channel (tetrahedral) microphone array; input features are the STFT phase component; position set 1, frame 0.

The mean absolute errors (MAE) of the position, azimuth, elevation, and distance from the sound source are used as a metric to evaluate the performance of the proposed method. For a single sound source, the MAE of the source position estimation is considered to be the Euclidean distance between the ground truth source position and the estimate of the source position e_i. The MAE is calculated using equation (2):

$$MAE = \frac{1}{N_{\text{Smp.}}} \sum_{i \in N_{\text{Smp.}}} e_i, \tag{2}$$

where $N_{\text{Smp.}}$ is the number of samples in the training data set.

For two-source position estimation, the association between the ground truth and the estimated source positions is unknown: the source position estimates are not ordered, and thus it is impossible to determine which estimate corresponds to which ground truth position. It was considered that for the MAE source position estimation evaluation, the estimated and the ground truth source positions will be paired based on the sum prediction error of both estimate-ground truth pairs, whichever combination (ground truth – estimate: 1-1, 2-2, or 1-2, 2-1) gives the least sum distance. This approach was used for both Cartesian and polar MAE evaluations for a two-source scenario.

To allow an objective comparison of our proposed method, a further evaluation of the proposed method was performed using the open access TAU Spatial Sound Events 2019 data sets that comprise a subset of tetrahedral microphone array signals [74], [75], [76] and compared with the baseline method provided along with the data sets [43] for the DCASE 2019 Challenge. The performance of our proposed method and the baseline method was evaluated using two frame-wise metrics: DoA error $\mathbf{DOA}_e$ (Equation (3)) and frame recall R_f (Equation (4)), described in [57]. For a recording of length T time frames, let $\mathbf{DOA}_R^t$ be the list of all reference DoAs in time frame t and $\mathbf{DOA}_E^t$ be the list of all estimated DoAs. The DoA error is defined as

$$\mathbf{DOA}_e = \frac{1}{\sum_{t=1}^{T} D_E^t} \sum_{t=1}^{T} \mathcal{H}(\mathbf{DOA}_R^t, \mathbf{DOA}_E^t), \tag{3}$$

where D_E^t is the number of DoAs in $\mathbf{DOA}_E^t$ in the t-th frame and $\mathcal{H}$ is the Hungarian algorithm to solve the assignment problem.

$$R_f = \frac{\sum_{t=1}^{T} 1(D_R^t = D_E^t)}{T}, \tag{4}$$

where D_R^t is the number of DoAs in $\mathbf{DOA}_R^t$ in the t-th frame, $1()$ is the indicator function that produces an output if the $(D_R^t = D_E^t)$ condition is met, otherwise returns zero.

III. RESULTS

After evaluating all trained CNN architectures, the MAE for source position estimation was calculated between the estimated source(s) position(s) and the ground truth source(s) positions(s). The results are provided in Table 2.

Furthermore, the results are presented in Figures 6, 7 and 8, respectively, for the estimation of the position of the single source from the maximum of the 3D matrix, the estimation of the position of the single source from the thresholding of the 3D matrix and the clustering of K-means, and the estimation of the position of two sources from the thresholding of the 3D matrix and the clustering of K-means.

Additionally, source position estimation errors were separately evaluated in the polar coordinates for azimuth, elevation, and distance. This was done to investigate the performance of the proposed method when used for the source DoA estimation task. The results of azimuth estimation are presented in Table 3. The results of the elevation estimation are presented in Table 4. The results of distance estimation are presented in Table 5. Source position estimation errors in polar coordinates were calculated by converting the Cartesian

TABLE 2. Source position estimation MAE values at different Q and σ; minimum MAE highlighted.

		3D matrix maximum		K-means clustering			
		1 source				2 sources	
		Noise	Speech	Noise	Speech	Noise	Speech
Q, [m]	σ, [m]	MAE, [m]	MAE, [m]	MAE, [m]	MAE, [m]	MAE, [m]	MAE, [m]
0.25	0.25	2.51	2.60	0.79	0.94	2.74	2.74
0.25	0.5	1.26	1.39	**0.62**	**0.76**	1.10	1.09
0.25	1	**0.99**	**1.10**	0.81	0.91	1.18	1.17
0.5	0.25	2.18	2.32	0.67	0.86	1.19	1.17
0.5	0.5	2.29	2.35	0.69	0.82	**1.08**	**1.08**
0.5	1	1.05	1.14	0.84	0.94	1.18	1.18
1	0.25	2.73	2.73	0.97	1.10	1.41	1.40
1	0.5	1.92	2.00	0.81	0.91	1.17	1.16
1	1	1.11	1.22	0.89	0.99	1.20	1.20

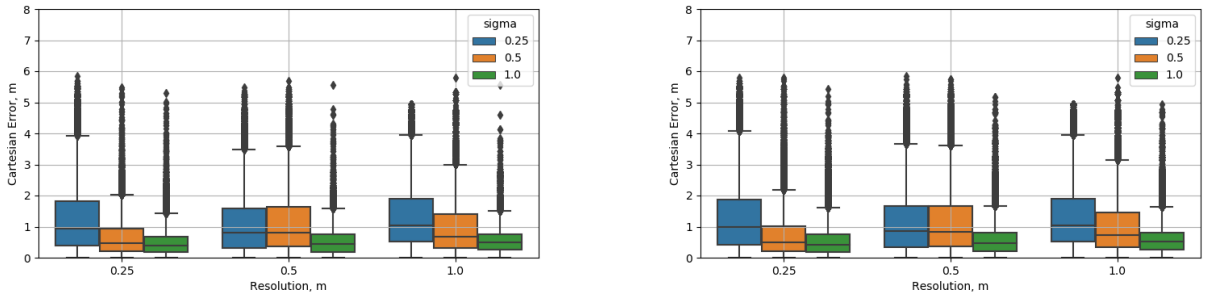


FIGURE 6. Sound source position estimation errors at different resolution and sigma values, noise (a) and speech (b) signals, 100 epochs, 1 source, coordinates obtained from 3D matrix maximum.

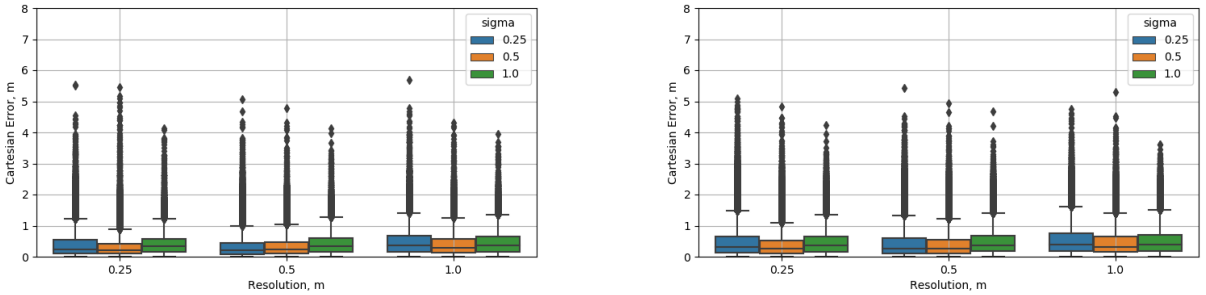


FIGURE 7. Sound source position estimation errors at different resolution and sigma values, noise (a) and speech (b) signals, 100 epochs, 1 source, coordinates obtained via K-means clustering.

source coordinates to polar coordinates, with the center of the microphone array considered as the origin point of the polar coordinate system. The azimuth and elevation errors are expressed in degrees.

The results of the evaluation of the source azimuth prediction errors are presented in Fig. 9 for single source localization when the source coordinates are obtained from the maximum of the 3D matrix, Fig. 10 for single source localization when the source coordinates are obtained via K-means clustering, and Fig. 11 for two-source localization when the source coordinates are obtained via K-means clustering.

An illustration of the 3D ground truth matrix with a single active source ($Q = 0.25$ m, $\sigma = 1$ m) and the corresponding 3D matrix predicted by CNN for input features with noise and speech sources are shown in Fig. 12. The center of the Gaussian blob in the ground-truth 3D matrix corresponds to the position of the sound source. In the predicted 3D matrix, the coordinates of the element with the maximum value, converted to metric coordinates using the Q factor, are considered the estimated source coordinates (in the case of the 3D matrix maximum coordinate acquisition method).

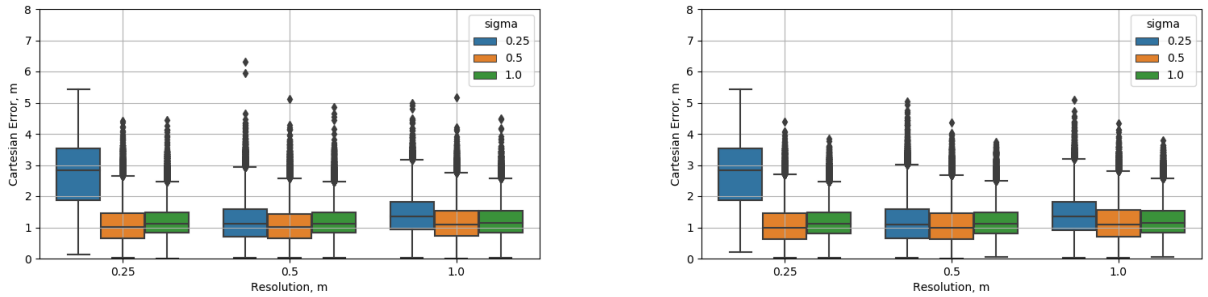


FIGURE 8. Sound source position estimation errors at different resolution and sigma values, noise (a) and speech (b) signals, 100 epochs, 2 sources, coordinates obtained via K-means clustering.

TABLE 3. Source azimuth estimation MAE values at different Q and σ; minimum MAE highlighted.

		Azimuth MAE, [°]					
		3D matrix maximum		K-means clustering			
		1 source				2 sources	
Q, [m]	σ, [m]	noise	speech	noise	speech	noise	speech
0.25	0.25	23.27	23.92	7.94	10.87	**18.97**	**18.95**
0.25	0.5	17.61	20.25	**5.88**	**8.48**	20.51	20.97
0.25	1	**12.59**	**15.71**	9.85	12.23	19.94	20.13
0.5	0.25	20.89	22.12	6.94	10.16	21.46	21.87
0.5	0.5	19.92	21.07	7.01	9.98	20.44	20.95
0.5	1	14.50	16.50	10.33	13.05	20.33	20.64
1	0.25	24.74	24.37	11.41	14.12	22.91	23.55
1	0.5	19.52	20.89	9.33	11.75	21.31	22.00
1	1	14.93	17.46	11.97	14.65	20.82	21.09

TABLE 4. Source elevation estimation MAE values at different Q and σ; the minimum MAE is highlighted.

		Elevation MAE, [°]					
		3D matrix maximum		K-means clustering			
		1 source				2 sources	
Q, [m]	σ, [m]	noise	speech	noise	speech	noise	speech
0.25	0.25	60.72	65.24	5.77	8.63	73.01	72.98
0.25	0.5	9.48	13.08	4.02	**5.29**	49.45	48.00
0.25	1	**6.49**	**9.39**	4.57	5.76	49.21	47.94
0.5	0.25	46.68	52.72	4.85	7.05	54.29	52.49
0.5	0.5	57.02	61.01	**3.91**	5.31	48.63	**47.70**
0.5	1	7.21	9.95	4.65	5.75	**48.49**	47.68
1	0.25	78.58	78.55	8.75	12.06	59.34	58.19
1	0.5	38.37	42.62	4.79	6.49	50.09	48.43
1	1	9.95	12.96	4.51	5.66	49.14	48.17

An illustration of the estimation of the position of two speech sources through the thresholding of the 3D matrix and the clustering of K-means is illustrated in Fig. 13. The resolution of the matrix is $Q = 0.5$ m and $\sigma = 0.5$ m. The clustering method is available for single-source and multiple-source scenarios. In Figures 13c and 13f estimation results are shown with one speech source replaced by noise.

When evaluating the proposed method using real-world tetrahedral microphone array signals, CNN was trained on a synthetic data set that was created in a simulated enclosure with geometry and acoustic parameters that corresponded to those of the real enclosure. A synthetic testing data set was created to match the positions and source signals from the real-world data set, as well as for evaluation purposes,

TABLE 5. Source distance estimation MAE values at different Q and σ; minimum MAE highlighted.

		Distance MAE, [m]					
		3D matrix maximum		K-means clustering			
		1 source				2 sources	
Q, [m]	σ, [m]	noise	speech	noise	speech	noise	speech
0.25	0.25	0.92	0.93	0.59	0.68	1.26	1.26
0.25	0.5	0.80	0.82	**0.48**	**0.59**	0.98	**1.00**
0.25	1	**0.69**	**0.71**	0.61	0.67	**1.02**	1.02
0.5	0.25	0.86	0.86	0.52	0.64	**1.02**	1.02
0.5	0.5	0.85	0.87	0.53	0.61	0.98	0.99
0.5	1	**0.69**	0.72	0.63	0.68	1.03	1.03
1	0.25	1.10	1.09	0.69	0.74	1.06	1.04
1	0.5	0.84	0.87	0.61	0.67	**1.02**	1.01
1	1	0.72	0.76	0.64	0.69	**1.02**	1.01

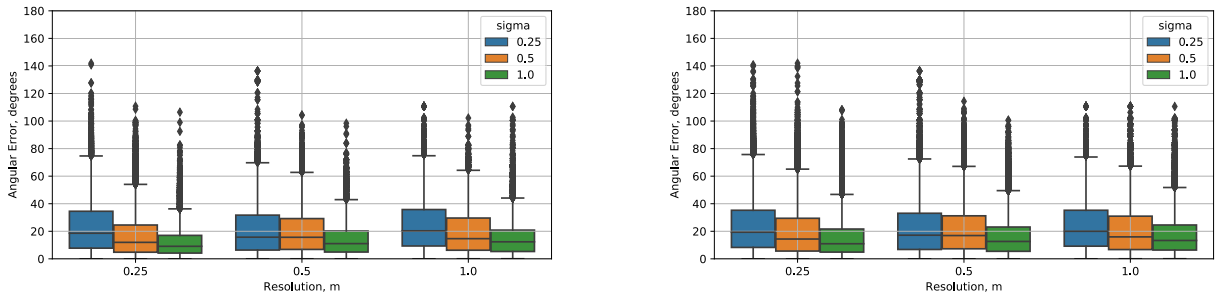


FIGURE 9. Sound source azimuth estimation errors at different resolution and sigma values, noise (a) and speech (b) signals, 100 epochs, 1 source, coordinates obtained from 3D matrix maximum.

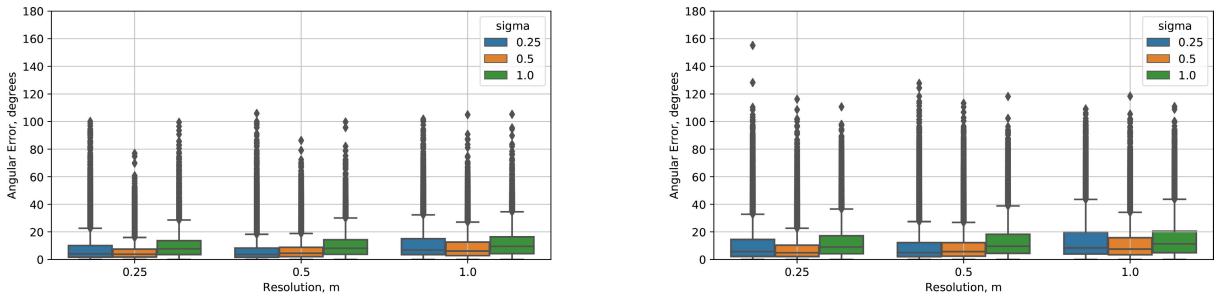


FIGURE 10. Sound source azimuth estimation errors at different resolution and sigma values, noise (a) and speech (b) signals, 100 epochs, 1 source, coordinates obtained via K-means clustering.

using a method described earlier. The output matrix values $Q = 0.25$ m and $\sigma = 0.5$ m at which the lowest single speech source position estimation errors were obtained, since the real world source signals were also speech signals. The mean absolute localization error for a single speech source using the synthetic data set was 0.9 m. Using the signals of the real-world microphone array, the mean absolute localization error of a single speech source was 2.46 m. This indicates that while the proposed CNN trained on simulated signals is not able to localize the speech source using real microphone signals with the same accuracy.

The results of the evaluation of our proposed method using the TAU Spatial Sound Events 2019 - Microphone Array data set are presented in Table 6, together with the results of the evaluation of the baseline method. The DoA error of the proposed method was 18.2°, while the baseline method showed an error of 38.1°. The frame recall of the proposed method was 63%, and the baseline method showed 83.4%.

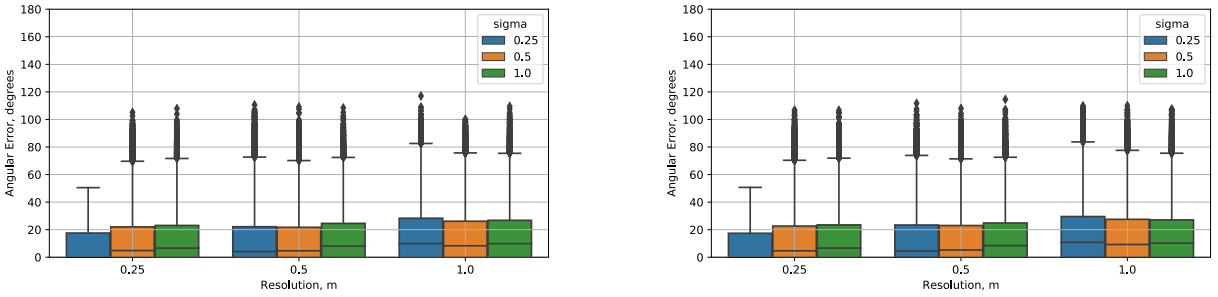


FIGURE 11. Sound source azimuth estimation errors at different resolution and sigma values, noise (a) and speech (b) signals, 100 epochs, 2 sources, coordinates obtained via K-means clustering.

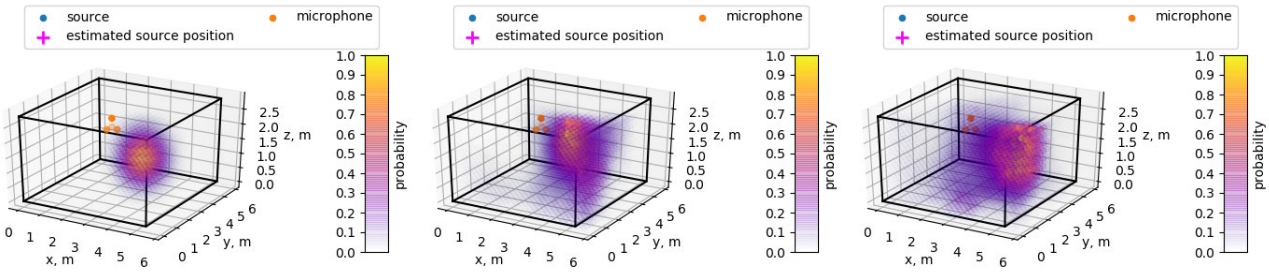


FIGURE 12. Examples of a ground truth 3D matrix (a) with one sound source ($Q = 0.25$ m, $\sigma = 1$ m) and CNN estimated 3D matrix for input features with respectively, a single noise (b) and a single speech (c) source; point color and opacity is proportional to the value of the matrix element; blue point show the ground truth position of the sound source; magenta cross show the estimated coordinates of the sound source (the position of estimated 3D matrix element with maximum value).

TABLE 6. Evaluation metric scores for the proposed method and the baseline method for the TAU Spatial Sound Events 2019 - Microphone Array data set.

	DoA Error, [°]	Frame Recall, [%]
Proposed Method	18.2	63.0
DCASE2019_MIC_baseline [43]	38.1	83.4

Furthermore, when evaluating the proposed method using the data set, the MAE of the estimation of the source position in the Cartesian coordinate system was 0.87 m.

A. DISCUSSION

In our work, we propose a novel method for the 3D localization of sound sources. The proposed method solves a classification problem with an uncertain number of sound sources, as opposed to a regression problem, where the number of sound sources must be known a priori. Current state-of-the-art methods successfully solve a regression problem in three dimensions or a classification problem in two dimensions. A recent survey of deep learning methods for the localization of single and multiple sound sources is provided in the work of Grumiaux et al. [12]. A solution to classify objects in 3D is relatively new. Only a few research works have previously proposed this idea [18], [47]. Both works used Multi-Layer Perceptron architectures from two to eight hidden layers. The method proposed in this paper uses a three-dimensional matrix as CNN output, where each point in the 3D matrix is a point in space and is classified into a desired class.

In this paper, we show several solutions for 3D localization: using 3D matrix maximums and K-means clustering for 1 and 2 sound sources. It should be noted that the proposed method is a derivation of our previous work [68], [69]. Here, a CNN-based 2D DoA estimation algorithm for multiple acoustic sources was shown to outperform state-of-the-art methods (such as the implementation of SRP-PHAT [71]).

As can be seen in Table 6, the proposed method outperforms the baseline method [43]. DoA error was reduced by more than 52%. However, the frame recall of the proposed method was lower than the baseline, achieving 63% frame recall compared to the baseline score (83.4%). This indicates that the proposed method is able to estimate the DoA of the acoustic source more accurately than the baseline, although it does not estimate as accurately as the baseline whether the acoustic source is active within the frame of the microphone array signals. It can be speculated that this behavior is due to reverberation, where the acoustic signal is still present within the enclosure after the acoustic source becomes inactive.

As can be seen in Table 2, the lowest source position estimate for the localization of a single noise source using the maximum 3D matrix finding as the coordinate estimation method was 0.99 m with matrix resolution $Q = 0.25$ m

and $\sigma = 1$ m. This shows that the source position estimation error is bound to the resolution of the output matrix, with the finest resolution producing the least errors. For the matrix maximum finding algorithm, this is expected behavior, as the uncertainty of source position estimation is equal to half the resolution. For the Gaussian kernel spread, the lowest source position error estimates were obtained at the largest spread values. This indicates that the proposed smoothness representation of the acoustic characteristics actually helps the neural network to learn the mapping between the acoustic characteristics and the position of the source. For the location of a single speech source, the lowest MAE = 1.10 m was achieved at the same values Q and σ. This shows that the trained neural network was able to generalize to various types of broadband acoustic source signals. When using the K-means clustering source coordinate estimation method, the lowest MAE is achieved for both noise (MAE = 0.62 m) and speech (MAE = 0.76 m) signals are achieved at $Q = 0.25$ m and $\sigma = 0.5$ m, which is a 37% improvement for noise source localization and 31% improvement for speech source localization. This can be explained by the fact that the coordinates of the cluster centers are not bound to the resolution of the output grid, thus allowing one to estimate the source coordinates with less uncertainty than would otherwise have been imposed by the resolution of the output grid. The lowest source position estimation errors were obtained with narrower Gaussian kernel spreads. It can be speculated that smaller Gaussian kernels lead to more localized clusters of thresholded output matrix elements, thus resulting in lower uncertainty of a cluster center estimation.

For the localization of two sources, the smallest MAE = 1.08 m for both noise and speech sources were achieved at $Q = 0.5$ m and $\sigma = 0.5$ m. The lower resolution at which the lowest position estimation errors of the two sound sources were obtained corresponds to the lower number of spatial classes that the neural network needs to learn to map acoustic features. For multiple sound sources, a single acoustic feature can be assigned to multiple spatial classes. It can be speculated that the neural network is unable to accurately learn to map acoustic features to spatial classes when the number of spatial classes is very high.

As shown in Table 3, the MAE of the lowest source azimuth estimation for the single noise source localization using the maximum 3D matrix finding as the coordinate estimation method was 12.59° with matrix resolution $Q = 0.25$ m and $\sigma = 1$ m, which corresponds to the results of the Cartesian source localization. For single speech signal azimuth estimation, lowest MAE = 15.71° was achieved at the same values Q and σ. If the single source position estimation method was used based on K-means clustering, the lowest MAE estimate was obtained for both noise and speech sources at $Q = 0.25$ m and $\sigma = 0.5$ m. Source azimuth estimation MAE was 5.88° (53% improvement) for the noise source and 8.84° (46% improvement) for the speech source. For the estimation of the azimuth of 2 sources, the smallest MAE = 18.79° for noise source and MAE = 18.95° for speech source were achieved at $Q = 0.25$ m and $\sigma = 0.25$ m.

As can be seen in Table 4, the lowest source elevation estimate for the localization of a single noise source using the maximum 3D matrix finding as the coordinate estimation method was 6.49° with matrix resolution $Q = 0.25$ m and $\sigma = 1$ m, which corresponds to the Cartesian source localization results and the azimuth estimation results. For single speech signal azimuth estimation, lowest MAE = 9.39° was achieved at the same values Q and σ. When using the single source position estimation method based on clustering of K-means, the lowest source elevation estimation MAE = 3.91° (40% improvement) for the noise source was obtained at $Q = 0.5$ m and $\sigma = 0.5$ m. For the speech source, lowest MAE = 5.29° (44% improvement) was achieved at $Q = 0.25$ m and $\sigma = 0.5$ m. For the two source elevation estimation, smallest MAE = 48.49° for noise source was achieved at $Q = 0.5$ m and $\sigma = 1$ m and MAE = 47.70° for speech source was achieved at $Q = 0.5$ m and $\sigma = 0.5$ m.

As can be seen in Table 5, the lowest source distance estimate for the location of a single noise source using the maximum 3D matrix finding as the coordinate estimation method was 0.69 m with matrix resolution $Q = 0.25$ m and $Q = 0.5$ m, and $\sigma = 1$ m. For single speech signal azimuth estimation, the lowest MAE = 0.71 m was achieved at the $Q = 0.25$ m and $\sigma = 1$ m, which corresponds to the source localization results using Cartesian coordinate system and the azimuth estimation results. When the single source position estimation method based on clustering of K-means was used, the lowest MAE of the MAE of source distance estimation MAE = 0.48 m (30% improvement) for the noise source at $Q = 0.25$ m and $\sigma = 0.5$ m. For the speech source, the lowest MAE = 0.59 (17% improvement) was achieved at $Q = 0.25$ m and $\sigma = 0.5$ m. For the two source elevation estimation, the smallest source distance estimation MAE = 1.02 m for the noise source was achieved at all values Q and $\sigma = 0.25$ m or $\sigma = 1$ m. Lowest source distance estimation MAE = 1 m for speech source was achieved at $Q = 0.25$ m and $\sigma = 0.5$ m.

To compare the results of azimuth and elevation estimation with the three-dimensional position estimation of the source, it can be seen that for the single source localization scenario, the lowest azimuth and elevation estimation errors are obtained using the maximum matrix finding method at the same values Q and σ, confirming the reasoning expressed earlier. This also holds true for clustering-based single-source azimuth estimation. For the two-source localization scenario, the lowest azimuth estimation errors are obtained for the finest grid and the smallest Gaussian kernel spread. This shows that, while the finer grid resolution and smaller clusters lead to a better azimuth estimation, the distance estimation is not accurate in this case. It can also be stated that the elevation estimation of the acoustic sources is less accurate than the elevation estimation. This might be attributed to the fact that the microphone array has three microphones in the same horizontal plane, while only the fourth microphone is

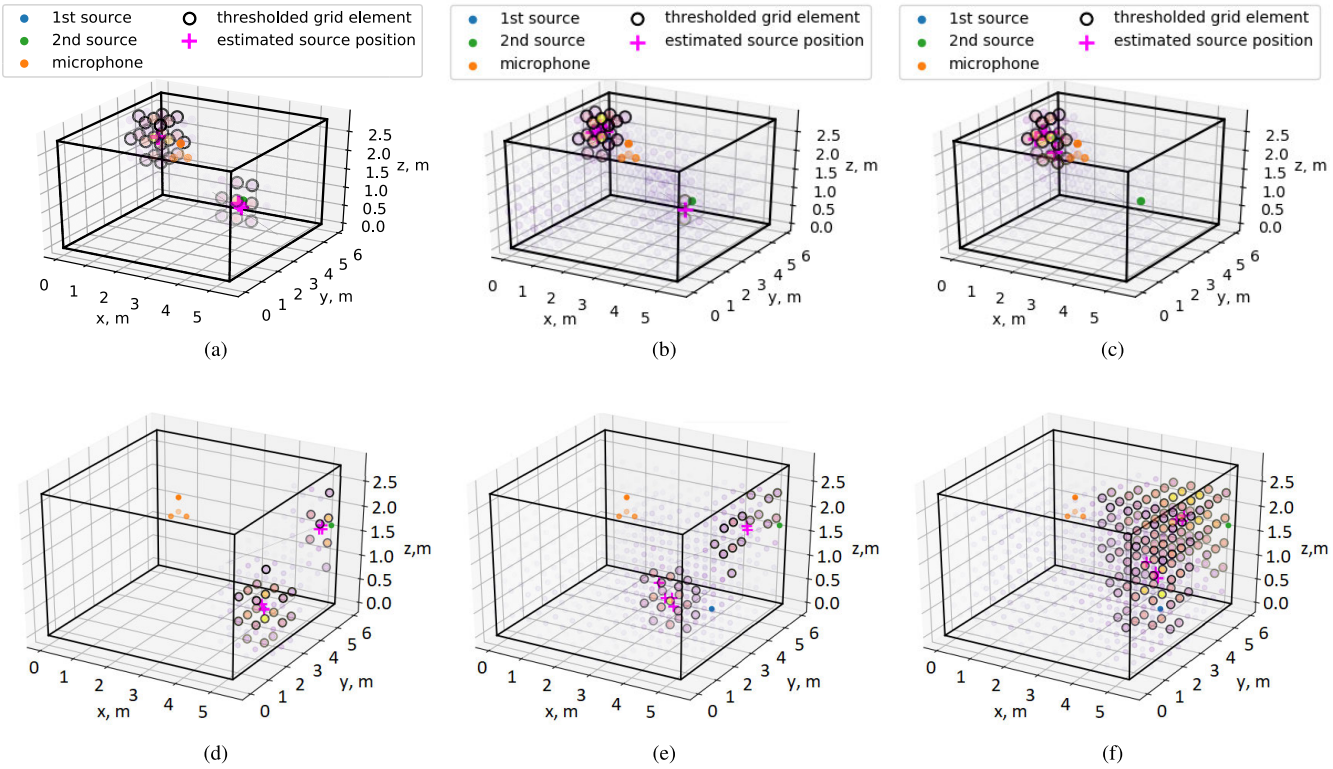


FIGURE 13. An example of a ground truth 3D matrix with two sources ($Q = 0.5$ m, $\sigma = 0.5$ m), when sources located in opposite sides of the microphone array (a) and when sources are located near the limits of the enclosure (d); CNN estimated 3D matrix for input feature with two speech sources (b, e); CNN estimated 3D matrix for input feature with a speech source and a noise source (c, f); point color and opacity is proportional to the value of the matrix element; points with black edges show the matrix elements remaining after the thresholding; blue and green points show the ground truth positions of the sound sources; magenta crosses show the estimated coordinates of the centers of clusters of the thresholded matrix elements; mean of the coordinates of these centers are the estimated source position.

vertically noncoplanar, leading to reduced information about the vertical positions of the sound sources contained within the acoustic features.

There are few limitations of the proposed method. In the scenario, when sound sources are located near the edge of the enclosure, the estimated source positions have higher errors due to the effects of the reverberant field, as shown in Fig. 13e. During the training, CNN learns to extract features from the STFT phase component and to map those extracted features to 3D matrix points with a particular probability. Thus, in the scenario where one sound source (out of two) is replaced by the noise source, the proposed method estimates two sound sources near the location of the remaining sound source, as shown in Fig. 13c. In the cases when the sound and noise sources are near, the output of the 3D matrix may have some elements remaining after thresholding (Fig. 13f). Clustering such output may lead to inaccurate results. The clustering-based sound source position estimation method, in theory, allows an arbitrary number of acoustic sources to be localized at any given moment. Additionally, if the number of sources is unknown, one can apply a source counting system in advance or set a threshold on this estimated probability, which implicitly provides source counting [77]. Furthermore, a carefully chosen automated clustering algorithm should remove this limitation. Nevertheless, in this investigation the CNN is only trained on target 3D matrixes with up to two clusters present. In future research, we plan to investigate the effects of automatic clustering algorithms and efficiency of subspace, density-based, and other non-hierarchical clustering methods.

IV. CONCLUSION

A method for sound source localization in a three-dimensional space using a tetrahedral microphone array and a CNN with STFT phase input features was proposed, and its performance was evaluated on a semi-synthetic audio data set.

Two methods were proposed for the estimation of the sound source positions from the three-dimensional output matrix: one based on finding the indices of the maximum valued element of the matrix, and another based on thresholding the 3D matrix element values and clustering of the remaining elements, with the center coordinates of the clusters considered the source position estimates.

An extensive experiment was carried out to investigate the influence of the resolution Q and Gaussian smoothing (spread parameter σ) on the accuracy of the sound source localization using the proposed method.

After discussing the results of the investigation of the proposed CNN-based 3D source position estimation method,

it can be concluded that it is possible to localize one or two sound sources within a 3D space using a CNN with STFT phase component of the tetrahedral microphone array signals as the input feature. Using the proposed clustering-based method instead of a 3D matrix maximum value element coordinate finding-based method provides at least 31% lower MAE for the estimation of the position of the source. Using the proposed method, it is possible to estimate the 3D coordinates of two simultaneously active sound sources with a position estimation MAE as low as 1.08 m for both noise and speech sources. It is possible to estimate DoAs of two simultaneously active speech or noise sources with azimuthal MAE as low as 18.97° and elevation MAE as low as 48.49°.

To further improve the method presented in this paper, it would be relevant to investigate alternative clustering methods for the CNN output matrix. Additionally, it could be investigated how the proposed model performs on more than two active sound sources, and its performance under variable acoustic conditions.

REFERENCES

[1] M. U. Liaquat, H. S. Munawar, A. Rahman, Z. Qadir, A. Z. Kouzani, and M. A. P. Mahmud, "Localization of sound sources: A systematic review," *Energies*, vol. 14, no. 13, p. 3910, Jun. 2021.

[2] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robot. Auton. Syst.*, vol. 96, pp. 184–210, Oct. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889016304742

[3] A. Toma, N. Cecchinato, C. Drioli, G. L. Foresti, and G. Ferrin, "CNN-based processing of radio frequency signals for augmenting acoustic source localization and enhancement in UAV security applications," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, May 2021, pp. 1–5.

[4] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," *J. Acoust. Soc. Amer.*, vol. 146, no. 5, pp. 3590–3628, 2019.

[5] W. Ma, H. Bao, C. Zhang, and X. Liu, "Beamforming of phased microphone array for rotating sound source localization," *J. Sound Vibrat.*, vol. 467, Feb. 2020, Art. no. 115064.

[6] N. Dey and A. S. Ashour, *Direction of Arrival Estimation and Localization of Multi-Speech Sources*. Berlin, Germany: Springer, 2018.

[7] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, vol. 1, Apr. 1997, pp. 375–378. [Online]. Available: http://ieeexplore.ieee.org/document/599651/

[8] A. D. Firoozabadi, P. Irarrazaval, P. Adasme, D. Zabala-Blanco, P. P. Játiva, and C. Azurdia-Meza, "3D multiple sound source localization by proposed T-shaped circular distributed microphone arrays in combination with GEVD and adaptive GCC-PHAT/ML algorithms," *Sensors*, vol. 22, no. 3, p. 1011, Jan. 2022.

[9] S. Y. Lee, J. Chang, and S. Lee, "Deep learning-enabled high-resolution and fast sound source localization in spherical microphone array system," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[10] M. Jia, J. Sun, and C. Bao, "Real-time multiple sound source localization and counting using a soundfield microphone," *J. Ambient Intell. Humanized Comput.*, vol. 8, no. 6, pp. 829–844, Nov. 2017.

[11] D. Su, T. Vidal-Calleja, and J. V. Miro, "Towards real-time 3D sound sources mapping with linear microphone arrays," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1662–1668.

[12] P.-A. Grumiaux, S. Kitic, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," 2021, *arXiv:2109.03465*.

[13] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 103–116, Apr. 2018.

[14] E. Vargas, J. R. Hopgood, K. Brown, and K. Subr, "On improved training of CNN for acoustic source localisation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 720–732, 2021.

[15] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 178–188, Jan. 2019.

[16] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Tokyo, Japan, Nov. 2013, pp. 2927–2932.

[17] N. Ma, G. J. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015, pp. 160–164.

[18] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, "On sound source localization of speech signals using deep neural networks," in *Fortschritte der Akustik. DAGA*, vol. 15. Berlin, Germany: German Acoustical Society (DEGA), 2015, pp. 1510–1513.

[19] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2016, pp. 603–609.

[20] Y. Huang, X. Wu, and T. Qu, "DNN-based sound source localization method with microphone array," in *Proc. Int. Conf. Inf., Electron. Commun. Eng.*, 2018, pp. 191–197.

[21] A. Zermini, Y. Yu, Y. Xu, M. Plumbley, and W. Wang, "Deep neural network based audio source separation," in *Proc. 11th IMA Int. Conf. Math. Signal Process.* Guildford, U.K.: University of Surrey, 2016, pp. 1–4.

[22] Y. Huang, X. Wu, and T. Qu, "A time-domain unsupervised learning based sound source localization method," in *Proc. IEEE 3rd Int. Conf. Inf. Commun. Signal Process. (ICICSP)*, Shanghai, China, Sep. 2020, pp. 26–32.

[23] Y. Wu, R. Ayyalasomayajula, M. J. Bianco, D. Bharadia, and P. Gerstoft, "SSLIDE: Sound source localization for indoors based on deep learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4680–4684.

[24] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*. New York, NY USA: Audio Engineering Society, 2015.

[25] 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.

[26] W. Ma and X. Liu, "Compression computational grid based on functional beamforming for acoustic source localization," *Appl. Acoust.*, vol. 134, pp. 75–87, May 2018.

[27] Y. Hao, A. Kuccuk, A. Ganguly, and I. Panahi, "Spectral fluxbased convolutional neural network architecture for speech source localization and its real-time implementation," *IEEE Access*, vol. 8, pp. 197047–197058, 2020.

[28] F. Hubner, W. Mack, and E. A. P. Habets, "Efficient training data generation for phase-based DOA estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 456–460.

[29] P.-A. Grumiaux, S. Kitic, L. Girin, and A. Guerin, "Improved feature extraction for CRNN-based multiple sound source localization," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Dublin, Ireland, Aug. 2021, pp. 231–235.

[30] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in CNN based multisource DOA estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1594–1608, 2021.

[31] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 6125–6129.

[32] Y. Cao, T. Iqbal, Q. Kong, M. B. Galindo, W. Wang, and M. D. Plumbley, "Two-stage sound event localization and detection using intensity vector and generalized cross-correlation," Detection Classification Acoustic Scenes Events (DCASE) Challenge, New York, NY, USA, Tech. Rep. DCASE2019-Cao-74, 2019.

[33] F. Grondin, J. Glass, I. Sobieraj, and M. D. Plumbley, "Sound event localization and detection using CRNN on pairs of microphones," DCASE, New York, NY, USA, Tech. Rep. DCASE2019-Park-33, 2019.

[34] S. Park, S. Suh, and Y. Jeong, "Sound event localization and detection with various loss functions," DCASE, Tokyo, Japan, Tech. Rep. DCASE2020-Park-89, 2020.

[35] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of CRNN models," DCASE, New York, NY, USA, Tech. Rep. DCASE2019-Kapka-26, 2019.

[36] Y. Kim and H. Ling, "Direction of arrival estimation of humans with a small sensor array using an artificial neural network," *Prog. Electromagn. Res. B*, vol. 27, pp. 127–149, 2011.

[37] M. Yasuda, Y. Koizumi, S. Saito, H. Uematsu, and K. Imoto, "Sound event localization based on sound intensity vector refined by DNN-based denoising and source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 651–655.

[38] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 74–79.

[39] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," 2021, *arXiv:2102.07955*.

[40] L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "CRNN-based joint azimuth and elevation localization with the ambisonics intensity vector," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 241–245.

[41] Y. Lin and Z. Wang, "A report on sound event localization and detection," DCASE, New York, NY, USA, Tech. Rep. DCASE2019-Lin-110, 2019.

[42] K. Noh, J.-H. Choi, D. Jeon, and J.-H. Chang, "Three-stage approach for sound event localization and detection," DCASE, New York, NY, USA, Tech. Rep. DCASE2019-Chang-81, 2019.

[43] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, Mar. 2019.

[44] H. Phan, L. Pham, P. Koch, N. Q. K. Duong, I. McLoughlin, and A. Mertins, "Audio event detection and localization with multitask regression network," DCASE, Tokyo, Japan, Tech. Rep. DCASE2020-Phan-117, 2020.

[45] R. Singla, S. Tiwari, and R. Sharma, "A sequential system for sound event detection and localization using CRNN," DCASE, Tokyo, Japan, Tech. Rep. DCASE2020-Single-56, 2020.

[46] F. Ronchini, D. Arteaga, and A. Pérez-López, "Sound event localization and detection based on CRNN using rectangular filters and channel rotation data augmentation," 2020, *arXiv:2010.06422*.

[47] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 405–409.

[48] H. W. Lollmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *Proc. IEEE 10th Sensor Array Multichannel Signal Process. Workshop (SAM)*, Jul. 2018, pp. 410–414.

[49] E. Guizzo, R. F. Gramaccioni, S. Jamili, C. Marinoni, E. Massaro, C. Medaglia, G. Nachira, L. Nucciarelli, L. Paglialunga, M. Pennese, S. Pepe, E. Rocchi, A. Uncini, and D. Comminiello, "L3DAS21 challenge: Machine learning for 3D audio signal processing," in *Proc. IEEE 31st Int. Workshop Mach. Learn. for Signal Process. (MLSP)*, Oct. 2021, pp. 1–6.

[50] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[51] S. Sivasankaran, E. Vincent, and D. Fohr, "Keyword based speaker localization: Localizing a target speaker in a multi-speaker environment," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 1–6.

[52] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1335–1345, Aug. 2019.

[53] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 2814–2818.

[54] 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 *Proc. IEEE 26th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Salerno, Italy, Sep. 2016, pp. 1–6.

[55] Z. Lu, "Sound event detection and localization based on CNN and LSTM," DCASE, New York, NY, USA, Tech. Rep. DCASE2020-Sampathkumar-41, 2019.

[56] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 300–311, 2021.

[57] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 1462–1466.

[58] J. Zhang, W. Ding, and L. He, "Data augmentation and priori knowledge-based regularization for sound event localization and detection," DCASE, New York, NY, USA, Tech. Rep. DCASE2019-He-97, 2019.

[59] C. Schymura, B. Bönninghoff, T. Ochiai, M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, and D. Kolossa, "PILOT: Introducing transformers for probabilistic sound event localization," 2021, *arXiv:2106.03903*.

[60] D. Suvorov, G. Dong, and R. Zhukov, "Deep residual network for sound source localization in the time domain," 2018, *arXiv:1808.06429*.

[61] J. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates," *Sensors*, vol. 18, no. 10, p. 3418, Oct. 2018.

[62] S. Chytas and G. Potamianos, "Hierarchical detection of sound events and their localization using convolutional neural networks with adaptive thresholds," in *Proc. Detection Classification Acoustic Scenes Events Workshop*, 2019, pp. 50–54.

[63] H. Pujol, E. Bavu, and A. Garcia, "Source localization in reverberant rooms using deep learning and microphone arrays," in *Proc. 23rd Int. Congr. Acoust.*, 2019, pp. 1–9.

[64] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 451–455.

[65] T. Jenrungrot, V. Jayaram, S. Seitz, and I. Kemelmacher-Shlizerman, "The cone of silence: Speech separation by localization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 20925–20938.

[66] H. Sundar, W. Wang, M. Sun, and C. Wang, "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 4642–4646.

[67] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 8, pp. 1393–1407, Aug. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7458213/

[68] S. Sakavicius and A. Serackis, "Estimation of sound source direction of arrival map using convolutional neural network and cross-correlation in frequency bands," in *Proc. Open Conf. Electr., Electron. Inf. Sci. (eStream)*, Apr. 2019, pp. 1–6.

[69] S. Sakavicius and A. Serackis, "Estimation of azimuth and elevation for multiple acoustic sources using tetrahedral microphone arrays and convolutional neural networks," *Electronics*, vol. 10, no. 21, p. 2585, Oct. 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/21/2585

[70] S. Sakavičius, "Improvement of learning-based methods for localization of multiple sound sources," Dept. Electron. Syst., Vilniaus Gedimino Technikos Universitetas, Vilnius, Lithuania, 2021.

[71] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 351–355.

[72] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, and J. Kadlec, "The AMI meeting corpus: A pre-announcement," in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.* Berlin, Germany: Springer, 2005, pp. 28–39.

[73] S. Sakavicius, "Dataset for evaluation of the performance of the methods of sound source localization algorithms using tetrahedral microphone arrays," *Mokslas-Lietuvos Ateitis*, vol. 12, pp. 1–8, Feb. 2020. [Online]. Available: https://journals.vgtu.lt/index.php/MLA/article/view/11462

[74] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2019, pp. 10–14. [Online]. Available: http://hdl.handle.net/2451/60746

[75] S. Adavanne, A. Politis, and T. Virtanen. (May 2019). *TAU Spatial Sound Events 2019-Ambisonic and Microphone Array, Evaluation Datasets*. [Online]. Available: https://zenodo.org/record/3377088

[76] S. Adavanne, A. Politis, and T. Virtanen. (Feb. 2019). *TAU Spatial Sound Events 2019-Ambisonic and Microphone Array, Development Datasets*. [Online]. Available: https://zenodo.org/record/2599196

[77] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Eng. Appl. Artif. Intell.*, vol. 110, Apr. 2022, Art. no. 104743.

SAULIUS SAKAVIČIUS (Member, IEEE) was born in Druskininkai, Lithuania, in March 1991. He received the bachelor's and master's degrees and the Ph.D. degree in electronics engineering from Vilnius Tech University, in 2014, 2016, and 2021, respectively. He prepared his doctoral dissertation, in December 2016. Currently, he is working with the Electronics Faculty as a Lecturer. His research interests include audio signal processing, machine learning, and the application of artificial neural networks.

ARTŪRAS SERACKIS (Senior Member, IEEE) was born in 1980. He received the M.Sc. and Ph.D. degrees from Vilnius Tech, in 2004 and 2008, respectively. He was an Associate Professor, in 2012, and a Professor, in 2017, at Vilnius Tech University. Since 2022, he has been the Head of the Electronic Systems Department. He is the author of more than 60 scientific articles, two textbooks, and a monograph. His current research interests include real-time image and signal processing, the development of intelligent systems, and the application of intelligent systems.

VYTAUTAS ABROMAVIČIUS (Member, IEEE) was born in 1988. He received the bachelor's degree in electronics engineering from Siauliai University, Siauliai, Lithuania, in 2011, and the master's and Ph.D. degrees in electronics engineering from Vilnius Tech University, Vilnius, Lithuania, in 2015 and 2019, respectively. His research focus is exploring ways to reduce symptoms of biosensor-based asthenopia from stereoscopic images, signal processing, and deep learning. Since 2015, he has been working as an Assistant Professor, and since 2022, as an Associate Professor with the Department of Electronic Systems, Vilnius Tech University.