Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

The Neural-SRP method for universal robust multi-source tracking

Eric Grinstein[1], Student Member, IEEE,
Christopher M. Hicks[2] Toon van Waterschoot[3], Member, IEEE
Mike Brookes[1], Life Member, IEEE, and Patrick A. Naylor[1], Fellow, IEEE

[1]Electrical and Electronic Engineering Department, Imperial College London, UK
[2]CEDAR Audio Ltd., Cambridge, UK
[3]Department of Electrical Engineering (ESAT), KU Leuven, Belgium

Corresponding author: Eric Grinstein (email: e.grinstein@imperial.ac.uk).

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 956369 and from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

ABSTRACT Neural networks have achieved state-of-the-art performance on the task of acoustic Direction-of-Arrival (DOA) estimation using microphone arrays. Neural models can be classified as end-to-end or hybrid, each class showing advantages and disadvantages. This work introduces Neural-SRP, an end-to-end neural network architecture for DOA estimation inspired by the classical Steered Response Power (SRP) method, which overcomes limitations of current neural models. We evaluate the architecture on multiple scenarios, namely, multi-source DOA tracking and single-source DOA tracking under the presence of directional and diffuse noise. The experiments demonstrate that our proposed method compares favourably in terms of computational and localization performance with established neural methods on various recorded and simulated benchmark datasets.

INDEX TERMS Deep Learning, Multi-source tracking, Direction-of-Arrival (DOA), Sound Source Localization (SSL)

DIRECTION-of-Arrival (DOA) estimation uses the signals from a microphone array to estimate the angular position of one or more active sound sources relative to the array. Applications include event detection [1]–[3], camera steering [4] and sound source separation [5]–[7]. Although many classical, signal processing based methods such as Multiple Signal Classification (MUSIC) [8], Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [9] and SRP [10], [11] have been extensively explored over the last decades, state-of-the-art localization performance is usually currently obtained using deep learning methods [12], where a neural network model is trained to estimate the location of the desired sources using a feature representation of the multi-channel microphone signals.

Neural DOA estimators can be classified according to their input features as *Time/Frequency (T/F)* or *hybrid*. T/F networks (e.g. DoaNet [13]) typically process features such as the multichannel Short Time Fourier Transform (STFT), Generalized Cross-Correlation with Phase Transform (GCC-PHAT) or the raw audio signal. A disadvantage of these networks is inflexibility to the microphone geometry, i.e., the number of microphones and respective positions of the array. This requires retraining for each array geometry, a cumbersome task which limits their off-the-shelf usage as a general tool. This also requires companies providing multiple array geometries within their line of products, such as voice assistants, to maintain multiple training pipelines. In contrast, current hybrid networks (e.g. Cross3D [14]) overcome this limitation by processing an input feature set that is independent of the number of microphone channels and their geometry, typically obtained using a classical signal processing DOA estimator such as the SRP method which will be described in Sec. A. A limitation of this approach is that it inherits the limitations of the underlying DOA estimator, such as an assumption of anechoic propagation and the lack of robustness to directional noise sources.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2023.3340057

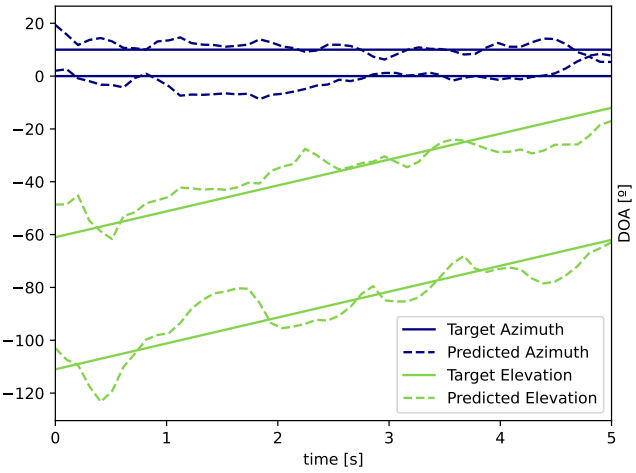

FIGURE 1. Example of Neural-SRP's output when tracking two moving sources. The panel shows the target and predicted azimuth and elevations.

Model	Causal	Universal	Multi-source
DOANet [13]			✓
Cross3D [14]	✓	✓	
Neural-SRP	✓	✓	✓

TABLE 1. Functional comparison of the proposed model and baselines. 'Universal' refers to the method's capacity of working on any microphone array geometry.

The main contribution of this work is Neural-SRP, a T/F neural localization method which overcomes the limitations of previous models. Tab. 1 shows a qualitative comparison of Neural-SRP with respect to Cross3D [14] and DOANet [13], arguably the literature's most established single and multi-source DOA estimation models. Unlike the DOANet, Neural-SRP is causal, therefore applicable to real-time applications, and universal, therefore applicable to arbitrary microphone geometries. In addition, unlike the Cross3D method, Neural-SRP is able to localize multiple sources simultaneously, as illustrated in Fig. 1. Finally, The proposed network is significantly smaller than the baselines. Code for the Neural-SRP architecture that can reproduce the experiments in this paper is available on Github [1].

Geometric independence is achieved by the introduction of two concepts, *pairwise processing* and *metadata fusion*. The former is inspired by the conventional SRP method, where a local feature is extracted between all microphone pairs, such local features then being summed to create a global feature. By providing the network with the microphone positions using a *metadata fusion* procedure, it is able to produce an *encoded pairwise spatial likelihood map*. After summation, the global feature is then decoded to estimate the sources' locations.

This paper continues as follows. Sec. I presents the signal model which will be used throughout this work. Sec. II presents a literature review of relevant neural methods for SSL, followed by a description of the conventional SRP method, from which our model takes inspiration. Sec. III describes our proposed model, followed by our experimental validation in Sec. IV. The results are discussed in Sec. V.

I. Problem definition and Signal Model

We define a 3-dimensional Cartesian system of coordinates centred at the position of a microphone array containing M microphones, whose known positions at discrete time index t are $\mathbf{v}_m(t) = [v_m^x(t) v_m^y(t) v_m^z(t)]^T$ for $1 \leq m \leq M$ at discrete time t. The goal of a DOA estimator is to provide an estimate of the set of positions $\mathcal{U}(t) = \{\mathbf{u}_1(t) ... \mathbf{u}_N(t)\}$, where $\mathbf{u}_n$ is defined analogously to $\mathbf{v}_m$, of the N active sound sources at time t. Each microphone m receives a signal *frame* of length L

$$\mathbf{x}_m(t) = \sum_{n=1}^{N} \mathbf{h}_{nm}(t) * \mathbf{s}_n(t) + \epsilon_m(t), \tag{1}$$

where the convolution operator is represented by $*$, $\mathbf{h}_{nm}(t) \in \mathbb{R}^R$ is the Room Impulse Response (RIR) vector of length R between source n and microphone m at time t, and $\mathbf{s}_n(t)$ is the signal frame emitted by source n at time t. In the case of Gaussian sensor noise, $\epsilon_m(t) \sim \mathcal{N}(\mathbf{0}, \sigma_m^2 I)$, where σ_m controls the Signal-to-Noise Ratio (SNR). In the case of a directional noise source, such as a fan, the noise term is defined as

$$\epsilon_m(t) = \sigma_m \big(\mathbf{h}_{m\epsilon} * \varepsilon(t)\big), \tag{2}$$

the impulse response $\mathbf{h}_{m\epsilon}$ convolved with a random signal $\varepsilon \sim \mathcal{N}(\mathbf{0}, I)$ scaled by a factor σ_m. Note the noise impulse response is not time-dependent, as we assume directional noise sources to remain spatially stationary at unknown position $\mathbf{u}_\epsilon$.

Although the sources can be located anywhere in the room, we are interested in their DOA, which we represent as a point on the unit sphere, i.e. $\|\mathbf{u}_n(t)\| = 1$. DOAs are also often represented as two angles, namely, *azimuth* and *elevation*. The azimuth is the angle between the x axis and the projection of $\mathbf{u}_n(t)$ in the horizontal xy plane, whereas the elevation is the angle between $\mathbf{u}_n(t)$ and the xy plane itself.

II. Prior art

Many approaches have been developed for the task of DOA estimation in the last decades. Arguably, the most established signal processing-based approach is the Steered Response Power (SRP) method, which was shown to be applicable to realistic scenarios containing noise and reverberation [15]. On the other hand, neural approaches have achieved state of the art performance at the cost of higher computation and limited generizability to unseen scenarios, a limitation which is overcome by our proposed method. The following sections provide a review of the SRP method and neural network approaches for DOA estimation.

[1] https://github.com/egrinstein/neural_srp

A. Steered Response Power

The main idea behind the SRP method [10], [11] is to map the temporal cross-correlation between a pair of microphone signal frames $(\mathbf{x}_i(t), \mathbf{x}_j(t))$, as well as their associated microphone positions into a Spatial Likelihood Function (SLF) [16] which associates a value $\text{SRP}_{ij}(\mathbf{p})$ for each candidate location $\mathbf{p} = [p_x\, p_y\, p_z]^T$ that is maximized at the true source locations. Note that the index t is omitted hereafter for conciseness. The *pairwise SRP* for a candidate location $\mathbf{p}$ is defined as [10], [11]

$$\text{SRP}_{ij}(\mathbf{p}; \mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \star \mathbf{x}_j)(\tau_{ij}(\mathbf{p})), \tag{3}$$

where the cross-correlation, represented by $\star$, between frames $\mathbf{x}_i$ and $\mathbf{x}_j$ is evaluated at the theoretical Time-Difference-of-Arrival (TDOA)

$$\tau_{ij}(\mathbf{p}) = \frac{f_s}{c}(\|\mathbf{p}_i - \mathbf{p}\| - \|\mathbf{p}_j - \mathbf{p}\|), \tag{4}$$

the difference in samples between the microphones located at $\mathbf{p}_i$ and $\mathbf{p}_j$ and the source $\mathbf{p}$. The speed of sound is c and f_s is the system's sampling frequency. In practice, GCC-PHAT [17] is commonly used instead of classical temporal cross-correlation. Finally, the global SRP is defined as the sum of all SRP pairs,

$$\text{SRP}(\mathbf{p}; \{\mathbf{x}_1, ..., \mathbf{x}_M\}) = \sum_{i=1}^{M} \sum_{j=i+1}^{M} \text{SRP}_{ij}(\mathbf{p}; \mathbf{x}_i, \mathbf{x}_j). \tag{5}$$

This represents the likelihood of a source being located at a candidate point $\mathbf{p}$, and the source location is estimated as

$$\hat{\mathbf{p}} = \underset{\mathbf{p}}{\text{argmax}} \text{SRP}(\mathbf{p}). \tag{6}$$

Other functions more flexible than the peak-picking in (6) can be used for the case of multiple active sources, such as peak subtraction [18] or sparsity-based [19] techniques. Lower computational cost can be achieved through usage of volumetric SRP variations [20], [21]. Also, the robustness of SRP can be improved in the case of moving sources/microphones by the inclusion of tracking algorithms [22], [23]. However, due to its formulation, the SRP method may exhibit multiple peaks in reverberant environments or in the presence of directional sources, as can be seen in Fig. 2.

B. Neural Networks for SSL

Neural networks have been widely applied for the task of DOA estimation using a centralized microphone array [12]. Multiple architectures have been proposed, including Convolutional Neural Networks (CNNs) [24], Multi-layer Perceptrons (MLPs) [25] or Convolutional Recurrent Neural Networks (CRNNs) [13]. They can also be classified by their output strategy, namely, regression or classification [26]. Finally, networks can be classified according to the input feature used, such as the complex-valued multichannel STFT [27], its phase [24], or the GCC-PHAT between all microphone pairs [13], [25]. If the input feature consists of the output of a classical signal processing method, such as

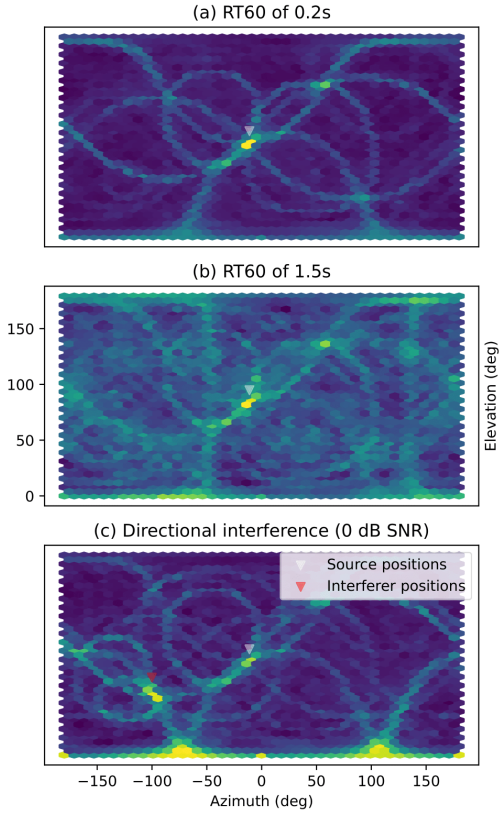

FIGURE 2. SRP maps generated for a simulated cuboid room containing one microphone array in its centre, as well as a source. (a) ideal, (b) reverberant and (c) noisy scenarios. The arrows point to the true source and interferer locations.

the SRP maps shown in Fig. 2, the network we classify it as *hybrid*. Otherwise, we shall classify it as *T/F*.

In [28] the concept of a dual-input neural network capable of jointly processing signals and metadata, such as the microphone positions, room dimensions and reverberation time for the task of positional SSL was introduced, allowing a T/F neural model to operate on distributed microphone arrays of unseen geometries, but with a fixed number of microphones. This constraint is removed in [29], where a spatial approach involving Graph Neural Networks (GNNs) is applied to the enhancement of SRP maps. In [30], an initial version of the Neural-SRP method is introduced for single-source positional SSL, where a network is trained to generate a likelihood map for each microphone pair. This work extends [29], [30] to the task of multi-DOA tracking. The remainder of this section focuses on the Cross3D [14] and DOANet [13] methods, which are respectively state-of-the-art hybrid and T/F models which serve as comparison baselines to our work.

The Cross3D method was proposed by Diaz-Guerra *et al.* [14] for the application of single-source DOA tracking. Their method can be interpreted as an image processing network, where its input is the 2D power map produced by the SRP method. The model's name is due to its architecture being a 3-dimensional causal CNN, where the three dimensions are azimuth, elevation and time. The authors show that the

model can be trained on simulated data generated using the image source method [31] and tested on a realistic dataset of real recordings. Recent work by the authors modified the approach to use icosahedral networks [32], significantly reducing the computational cost of Cross3D. Multi-source capabilities were also recently introduced in [33].

The DOANet method was proposed by Adavanne *et al.* [13] for tracking up to two simultaneous sound events. The main model used is a bidirectional CRNN [34]. The authors show that including tracking metrics defined in [35] significantly improved the model's performance. The output of the network consists of a vector of size 8, where the first 6 elements refer to the estimated source positions, and the last two represent the activity of each track, similar to a Voice Activity Detector (VAD).

III. Neural-SRP

A. Input Feature Set

The input feature of Neural-SRP consists of the GCC-PHAT of all pairs of microphone signal frames $(\mathbf{x}_i, \mathbf{x}_j)$, defined as

$$\mathbf{g}_{ij} = \mathrm{IDFT}\left(\frac{\boldsymbol{x}_i}{|\boldsymbol{x}_i|} \odot \frac{\boldsymbol{x}_j^*}{|\boldsymbol{x}_j|} \right), \tag{7}$$

the L-sized Inverse Discrete Fourier Transform (IDFT) of the element-wise product of the normalized frequency-domain frames $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where $\boldsymbol{x}_k = \mathrm{DFT}(\mathbf{x}_k)$ and $|\boldsymbol{x}_k|$ is the element-wise magnitude. The input feature consists of the GCC-PHAT between all microphone pairs, thus generating an input of shape $(M(M-1)/2, T, G)$, where T is the number of time-frames and G is the number of central GCC delays used. This selection has the advantage of reducing the input size and removing delays which are bigger than the maximum theoretical TDOA for the microphone array, computed as

$$G = 2 \max \left\{ \frac{\|\mathbf{v}_i - \mathbf{v}_j\| f_s}{c} \right\} + 2G_0 \tag{8}$$

where $1 \leq i < j \leq M$ and $G_0 \geq 0$ is a parameter to increase the feature size to values beyond the maximum theoretical TDOA, which increases performance in practice [13]. This input feature is also used by the DOANet model. However, while the DOANet model jointly processed all input features using a single network, our proposed model processes each pairwise feature independently to create a summable encoded likelihood map, allowing the network to accept any number of microphone pairs as its input.

B. Architecture

The Neural-SRP network is divided into two sub-networks, namely, a pairwise network $\mathscr{P}$ and a global decoder $\mathscr{D}$. The architecture is shown in Fig. 3 and is summarized as

$$\hat{\mathcal{U}} = \mathscr{D}\left(\sum_{i=1}^{M} \sum_{j=i+1}^{M} \mathscr{P}(\mathbf{g}_{ij}, \mathbf{v}_i, \mathbf{v}_j) \right). \tag{9}$$

The goal of $\mathscr{P}$ is to create an encoded and summable spatial likelihood feature for each signal pair, using GCC $\mathbf{g}_{ij}$ along with its respective microphone coordinates $(\mathbf{v}_i, \mathbf{v}_j)$. These features are them summed together, creating a global feature which is then decoded by $\mathscr{D}$ to estimate a set of locations $\hat{\mathcal{U}}$. The proposed method's name derives from the structural similarity between (9) and (5).

The pairwise network consists of a modified Convolutional Recurrent Neural Network (CRNN) architecture. The parameters of the pairwise network are shared across all pairs. Each pairwise GCC is first processed by a sequence of 2D convolutional blocks. To maintain causality, the kernel size in the time dimension is set to 1 and no pooling is applied in that dimension. Unit strides were used on convolutional layers. The resulting feature of shape (T, C_c^0, C_c^1) is transformed into shape (T, C_c) by flattening the last two dimensions of size, C_c^0, the number of output kernels, and C_c^1, the number of GCC bins after pooling.

To improve tracking performance, the resulting feature is then processed by a one-directional Recurrent Neural Network (RNN) of type Gated Recurrent Unit (GRU) [36]. To produce a spatially-aware feature, the microphone coordinates of each microphone in the pair are concatenated to each channel, followed by transforming this feature into an encoded likelihood map of shape C_p through the application of another MLP. An interpretation of this step is 'steering' the feature produced by the RNN according to the direction of the segment connecting the microphone pair's positions. We refer to [28] for a detailed discussion on methods for incorporation of metadata, namely microphone position information, for the improvement of SSL methods.

The decoder $\mathscr{D}$ consists of two independent MLPs as in the DOANet model. The first is an activity detector similar to a multichannel VAD, while the second outputs the $\hat{N}$ estimated locations. These outputs are implicitly related, in the sense that if the n^{th} activity detector indicates no activity, the values of the n^{th} estimated DOA should be ignored.

C. Training

Both pairwise and global networks are jointly optimized using the network's output. In the following, we shall define the loss function for each temporal instant t and will therefore omit this index. We define $\mathbf{U} = [\mathbf{u}_1 \dots \mathbf{u}_{\hat{N}}]$ and $\hat{\mathbf{U}} = [\hat{\mathbf{u}}_1 \dots \hat{\mathbf{u}}_{\hat{N}}]$ as the target and output DOA matrices respectively, where each column is a unit vector representing a true or estimated DOA. We also define $\mathbf{z}$ and $\hat{\mathbf{z}}$, $\hat{N}$-dimensional binary vectors which refer to the target and output activities. In the case where only a single source exists, the loss function is defined as

$$\mathcal{L}(\mathbf{U}, \hat{\mathbf{U}}, \mathbf{z}, \hat{\mathbf{z}}) = \alpha z_1 \|\mathbf{u}_1 - \hat{\mathbf{u}}_1\| + \beta \mathrm{BCE}(z_1, \hat{z}_1) \tag{10}$$

where the first term is the *Euclidean localization error* between the true and estimated DOA, weighted by the true activity, so as to ignore silent frames. The Euclidean error is employed in favour of the more interpretable angular error as previous works [14], [26] found it to yield better training results. The weighting factors α and

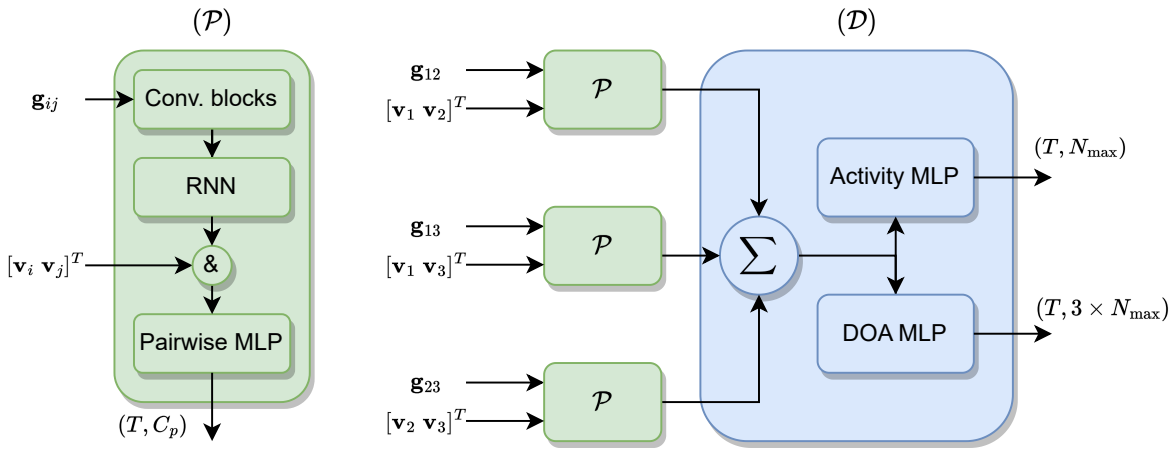

FIGURE 3. **Neural-SRP network architecture. Left, green: pairwise network $\mathcal{P}$. Right, blue: Global decoder $\mathcal{D}$, exemplified for a 3-microphone input. Symbol "&" represents concatenation.**

β are hyperparameters. The second term is the binary cross-entropy $\text{BCE}(z_1, \hat{z}_1) = (z_1 \log \hat{z}_1 + (1 - z_1) \log(1 - \hat{z}_1))$ between the true and target activity.

To prevent the loss function from diverging to $-\infty$, we clamp the maximum value of $\log(\cdot)$ to a constant B. When two or more sources are active, the training must take the *assignment problem* into account, so as not to penalize equivalent target and true permutations [37]. This problem can be defined as finding the association matrix $\mathbf{A}$, a permutation of the rows of the identity matrix of size $\hat{N}$. The optimal $\mathbf{A}$ minimizes the multi-source localization error, defined as

$$\mathcal{L}^{doa}(\mathbf{U}, \hat{\mathbf{U}}, \mathbf{z}) = |\mathbf{D} \odot \mathbf{A}| / |\mathbf{z}|, \tag{11}$$

where $[D]_{ij} = \|\mathbf{u}_i - \hat{\mathbf{u}}_j\|$ is the distance matrix between all target and output combinations, $\odot$ is the element-wise product, $|\cdot|$ is the matrix norm and $|\mathbf{z}|$ is the number of active sources. Although $\mathbf{A}$ can be deterministically computed using the Hungarian algorithm [38], the latter is not differentiable, hindering its application for training the neural network using a backpropagation procedure. We solve this problem in the same manner as [13], where a neural network is used to approximate the Hungarian method, and then used for training. The association matrix is also used for aligning the target and output activities $\mathbf{z}$ and $\hat{\mathbf{z}}$, after which the binary cross-entropy function is applied for each entry.

IV. Experimentation

Experiments were performed consisting of training/evaluating Neural-SRP and baselines on datasets of different complexities and characteristics, each serving the purpose of evaluating the method's performance in different conditions. Five datasets were used, three simulated and two recorded, which are described below. The Cross3D and DOANet baselines use the same architectural parameters and training procedures described in their respective original papers [13], [39].

The network parameters for Neural-SRP are summarized in Fig. 4, where the tensor output shapes are shown for each of the network's layers. Convolutional kernels of size $(3,3)$ were used on all convolutional layers. Max pooling with a kernel size 2 was applied to the GCC-PHAT dimension after all but the last convolutional layers. Parametric Rectified Linear Unit (PReLU) activation was used for all of the network's layers, apart from the RNN and DOA MLP output, which used a Hyperbolic Tangent (TANH) activation, and the activity output layer, which used sigmoidal activation. This architecture was chosen empirically. All the networks were implemented using the Pytorch library. The Adam optimizer was used for backpropagation.

A rectangular grid of size 64×32 was used for SRP, where the first dimension represents azimuth and the second elevation. The same configuration was used for generation of the input maps for the Cross3D baseline. The parameters used for the latter and the DOANet baseline were chosen similarly to those used in the respective original papers [13], [14].

A. Evaluation metrics

The main metric used for the single source experiment was the Root Mean Square Angular Error (RMSAE) [40], defined for a pair of positions $(\mathbf{u}, \hat{\mathbf{u}})$ each with azimuth and elevations (θ, ϕ) and $(\hat{\theta}, \hat{\phi})$ respectively, as

$$\mathcal{E}(\mathbf{p}, \hat{\mathbf{p}}) = \arccos^2(\cos\theta\cos\hat{\theta} + \sin\theta\sin\hat{\theta}\cos(\phi - \hat{\phi})), \tag{12}$$

where (12) was averaged for all frames in the dataset. For multiple sources, the localization error is defined for each correctly detected source using the ground truth association matrix $\mathbf{A}$. For the multi-source experiment, the detection metrics of precision, recall and the F1 score were also used, as defined in [13], [35]. These metrics are computed for each frame, based on the number of true and estimated sources $|\mathcal{U}|$ and $|\hat{\mathcal{U}}|$. $|\mathcal{U}|$ and $|\hat{\mathcal{U}}|$ are first used to compute the number true positive $\text{TP} = \min(|\mathcal{U}|, |\hat{\mathcal{U}}|)$,

false positive FP $= \max(0, |\hat{\mathcal{U}}| - |\mathcal{U}|)$ and false negative FN $= \max(0, |\mathcal{U}| - |\hat{\mathcal{U}}|)$ detections. Finally, the Precision (PR), Recall (RE) and F1 metrics are computed as:

$$\begin{aligned} \text{PR} &= \text{TP}/(\text{TP}+\text{FP}) \\ \text{RE} &= \text{TP}/(\text{TP}+\text{FN}) \\ \text{F1} &= 2(\text{PR}\times\text{RE})/(\text{PR}+\text{RE}). \end{aligned} \tag{13}$$

As in the single source experiments, the final metrics are obtained for the proposed method and baselines through averaging of all frame metrics in the dataset.

B. Datasets

The first three experiments were performed using simulated datasets, which we refer to as SimSW, SimDirect and SimRandMic. All datasets contain samples of a source moving in a 3-dimensional sinusoidal trajectory inside a cuboid-shaped reverberant room containing a compact stationary microphone array. The trajectories were generated by randomly selecting a start and end point inside the room, followed by randomly assigning a 3-dimensional vector referring to the frequency of oscillations within each direction. Finally, a second 3-dimensional vector is randomly generated representing the amplitude of each direction's oscillation. As in [14], the simulated datasets follow an "infinite-style" paradigm, meaning acoustic scenarios are randomly generated using the image source method [31] during training, i.e., no data is stored. The duration of each sample is 20 s. The ranges of the parameters are shown in Tab. 2. The sampling rate used for the simulations was equal to 16kHz.

Both the first and second datasets, named SimSW and SimDirect use the pseudo-spherical array geometry of the NAO Robot as described in the LOCATA dataset [41]. SimSW and SimDirect differ in the type of noise used, respectively, spatially white (SW) sensor noise and directional noise. The goal of these datasets is to assess the robustness of the algorithms to different types of noise. For the third dataset, named SimRandMic, a random array geometry was generated for each dataset sample. The goal of this dataset was to assess the methods' generalizability to unseen microphone geometries.

Parameter	Min. value	Max. value
RT60 (ms)	0.2	1
SNR (dB)	5	30
Oscillations	0	2
Oscill. amp. (m)	0	1
# mics. (SimMicRand)	4	12
Array radius (SimMicRand, cm)	5	10

TABLE 2. Parameter ranges for simulated datasets.

The datasets were generated using the gpuRIR Python library [39], which can simulate audio recordings of cuboid-shaped, reveberant rooms including an arbitrary number of moving sources and microphone arrays. Simulating moving sources/microphones is a computationally expensive task, as high quality scenes are typically rendered by generating one RIR using the Image Source method between each source-microphone pair at every few milliseconds, and auralizing audio signals by convolving the source signals and RIRs using the Overlap-Add strategy [39]. gpuRIR significantly reduces the computational time in comparison to other libraries such as Pyroomacoustics [42] by generating the RIRs in a Graphics Processing Unit (GPU).

In the SimSW and SimRandMic datasets, random Gaussian white noise is added to the auralized signals at the desired SNR, computed using (15). In the SimDirect dataset, a second source emitting random Gaussian noise is randomly placed at a static position inside the room. The auralized noise signal is added to the source signal by scaling it to the desired SNR, computed using the mean energy of both auralized signals across all frames. In the SimRandMic dataset, a spherical microphone array is generated for every sample first by uniformly sampling its radius and number of microphones from the values ranges shown in Tab. 2, followed by randomly placing the microphones on the sphere's boundary. An utterrance from the LibriSpeech dataset [43] is randomly chosen as source signal for each dataset sample. Each epoch consists of a network pass through all of the Librispeech dataset, although different scenes are generated for each epoch.

We define $\mathbf{y}_m(t) = \mathbf{x}_m(t) - \epsilon_m(t)$ as the idealized noiseless signal frame received at microphone m. We define the average array-wide power of all signal frames p_y as

$$p_y(t) = \frac{1}{LMT}\sum_{l=0}^{L-1}\sum_{m=1}^{M}\sum_{t=0}^{T-1} z(t)\mathbf{y}_m(t,l)^2, \tag{14}$$

where the ideal binary voice activity detector z is used to ignore silent frames. The array-wide power of each noise frame p_ϵ is defined analogously. We compute the array-wide *spatially white* SNR_{sw} as

$$\text{SNR}_{\text{sw}} = 10\log_{10}\frac{p_y}{p_\epsilon}. \tag{15}$$

The LOCATA dataset [41] was released as part of the 2018 IEEE AASP Challenge on acoustic source LOCalization And TrAcking. It consists of 6 tasks of increasing complexity. In this work, we select tasks 1, 3 and 5, namely, static, moving source and moving microphone localization. The dataset provides recordings from multiple microphone arrays. In this work, we only use recordings provenant from the NAO robot, which contains a pseudo-spherical 12-channel microphone array. The goal of this dataset is to assess the performance of the algorithms in a real environment, as well as their ability to generalize to a real environment through training on simulated data. The sampling rate of the dataset is 48 kHz.

The TAU-NIGENS Spatial Sound Events dataset [44] was originally released for the Sound Event Localization and Detection task of the 2021 Detection and Classification of Acoustic Scenes and Events (DCASE) challenge. It was

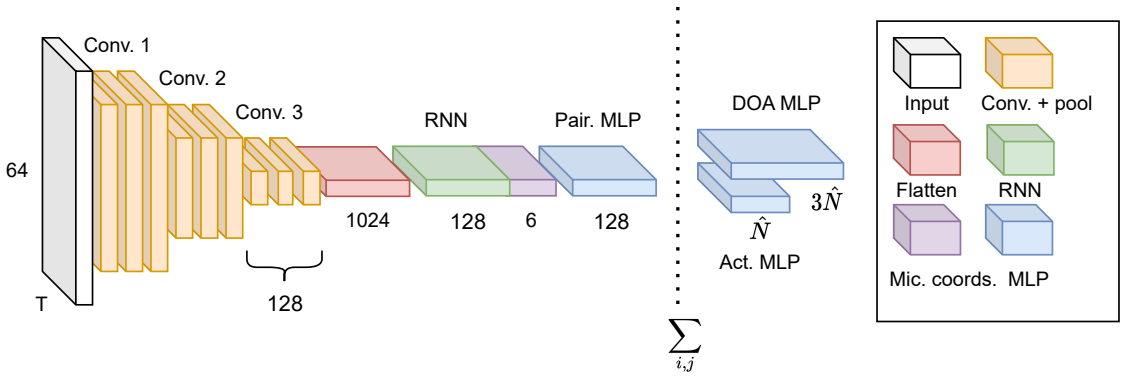

FIGURE 4. **Detailed view of Neural-SRP architecture, where the numbers show the output dimension of each layer. The dotted line separates the pairwise network $\mathcal{P}$ from the global decoder $\mathcal{D}$, which receives the sum of pairwise features as its input. The input layer consists of T frames of 64 central GCC-PHAT bins each. The mic. coords. layer is of shape $(T,6)$ where the three coordinates for each of the microphone in the pair are replicated for all frames.**

generated by filtering source signals from the NIGENS Sound Events database [45] using time-varying RIRs recorded on 13 different rooms of Tampere University, Finland. These RIRs were recorded using a 32-channel Eigenmike spherical microphone array and a Genelec G Three [2] loudspeaker. Instead of providing the full 32-channel recordings, equivalent compressed 4-channel tetrahedral signals are provided. The dataset is subdivided into 400 training, 100 validation and 100 testing 1-minute recordings of up to two simultaneous moving sources. The samples may be corrupted by directional, moving interference emitting signals belonging to a noise class from the NIGENS database. The goal of this dataset is to assess the performance of the Neural-SRP method for tracking multiple sources. The sampling rate of the dataset is 24 kHz.

C. Experiment 1: spatially white noise

In this experiment, we evaluate the performance of Cross3D, Neural-SRP and conventional SRP in the presence of independent White Gaussian Noise (WGN) added to each sensor. The neural models are trained for a duration of 80 epochs using a learning rate of 10^{-4}. As in [39], we use a frame size of 256 ms and a hop size of 192 ms. The noise signals are generated with unit variance, then scaled to the randomly selected SNR_{sw} by inverting (15), and then summed to the noiseless received signals. Both networks are trained using the range of SNRs defined in Tab. 2, and tested using a simulated dataset of unseen source signals from the Librispeech test set, as well as the unseen LOCATA dataset. The results are shown in Tab. 3. We also report the dependence of the localization error to reverberation and noise on the test dataset, as shown in Fig. 5.

D. Experiment 2: Directional noise

In this experiment, the Neural-SRP method is applied to the task of localizing a single speech source on a directional noise scenario, which is arguably more realistic than the diffuse case. For example, a directional noise source could be a fan, or a washing machine. The main difference from the experiment described in Sec. C is that, instead of adding independent noise to each microphone, the noise is itself modeled as a source in the room. In other words, for each training sample, the interferer is randomly placed within the room, with the restriction of being at least one meter away from the source and array. Then, a RIR between the microphones and interferer is computed, which is then convolved with a random unit variance Gaussian signal. Finally, the auralized result is scaled to the randomly assigned SNR in the same manner as (15). The results are shown in Tab. 3.

Model	SimSW.	LOCATA. (O)	SimDirect.	LOCATA. (D)
Cross3D	4.2	6.1	3.5	6.1
Neural-SRP	3.2	4.7	3.4	5.8
SRP	8.4	16.7	7.9	16.7

TABLE 3. **Average localization error for experiment 1 (first and second columns) and 2 (third and fourth columns). All values are expressed in degrees. LOCATA (O) and (D) are the results following training using SimSW and SimDirect respectively. Both entries in the aforementioned columns show the same value for SRP, as it is not trained.**

E. Experiment 3: Testing on an unseen geometry

To assess the proposed model's ability to generalize to unseen microphone geometries, we trained it using a dataset of multiple microphone array geometries, while testing it on the microphones mounted on the NAO robot head of the LOCATA dataset, a geometry which is unseen in the training dataset. Although the Cross3D method can be theoretically trained using variable microphone geometries, we were unable to train it using the SimRandMic dataset as the initialization of the SRP method was shown to be prohibitively costly, resulting in each epoch taking several hours on a GPU-enabled server. As a means of comparison, we use conventional SRP, as well as Neural-SRP trained

[2] https://www.genelec.com/g-three

using the SimSW dataset, i.e., a matched array geometry. The results can be seen in Tab. 4.

Model	Trained on	SimSW (°)	LOCATA (°)
Neural-SRP	SimRandMic	6.7	6.0
Neural-SRP	SimSW	6.7	3.4
SRP	N/A	7.9	16.7

TABLE 4. Average localization error for experiment 3

F. Experiment 4: Multi-source tracking

Model	Loc. err. (°)	Precision (%)	Recall (%)	F1 (%)
DOANet	9.4	88.6	81.9	85.1
Neural-SRP	8.2	92.0	79.9	85.4

TABLE 5. Average multi-source metrics and standard deviations of Neural-SRP and DOANet on the testing TAU-NIGENS dataset. Metrics and deviations were computed by averaging across the 3 training experiments.

In this experiment, the Neural-SRP method is applied to the task of multi-source tracking. We compared our method to the state-of-the-art DOANet model with parameters described in [13] on the TAU-NIGENS dataset. The network was trained three times for a duration of 80 epochs using a learning rate of 10^{-4}. The average localization error was computed on the validation dataset at the end of each epoch, and the network weights that obtained the lowest validation localization error were used for evaluating the unseen test set. The results are shown in Tab. 5, where each value is the average metric obtained for each training round. The metrics used were the localization error in degrees, for true positive matches, as well as classical tracking metrics, namely, precision, recall, and the F1 score, defined as a geometric average of the two aforementioned scores. An example output of Neural-SRP successfully tracking two simultaneous sources is shown in Fig. 1. As in [13], a frame of size 20 ms with a hop size of 10 ms was used.

G. Complexity comparison

In this section, the complexity of the proposed Neural-SRP model and baselines is presented in terms of number of parameters, computational time and number of Floating-Point Operations (FLOPS) for microphone array sizes $\{4, 8, 12\}$. The number of FLOPS is obtained through the use of the THOP Python library [3]. This library is not compatible with the SRP, so we compute the theoretical complexity of the latter theoretically, as in [46]. The inference time is measured as the clock difference taken for the model to produce an output for an input stimulus of duration of one-second. These results were obtained using a 16 GB Macbook Pro with an M1 chip and are shown in Tab. 6.

[3]https://github.com/Lyken17/pytorch-OpCounter

Model	# params (10^6)	Inf. Times (ms)	FLOPS (10^9)
Cross3D	5.62	398, 423, 450	20
DOANet	1.57, 1.59, 1.64	11, 20, 35	0.1, 0.2, 0.3
Neural-SRP	0.92	13, 75, 149	0.45, 2.1, 4.9
SRP	0	8, 25, 54	0.39, 1.8, 4.3

TABLE 6. Complexity analysis of Neural-SRP and baselines. The cells containing three numbers refer to 4, 8 and 12 microphones respectively.

V. Discussion and analysis

The single source experiments summarized in Tab. 3 show that Neural-SRP obtains favourable results in comparison to the Cross3D method both in the spatially white and directional noise scenarios, despite using a significantly smaller and more computationally efficient model. Like Cross3D, Neural-SRP can be trained using simulated data and tested using real recordings, as seen in the LOCATA results in Tab. 3. This is remarkable as, unlike Cross3D, Neural-SRP is required to learn its own spatial representation of sound. In other words, Neural-SRP is able to generalize despite having a less stringent inductive bias. Another relevant remark is that unlike SRP, Neural-SRP and Cross3D were able to eliminate the effect of a directional noise source, which is typically manifested as an additional peak in the GCC-PHAT (and therefore SRP) features.

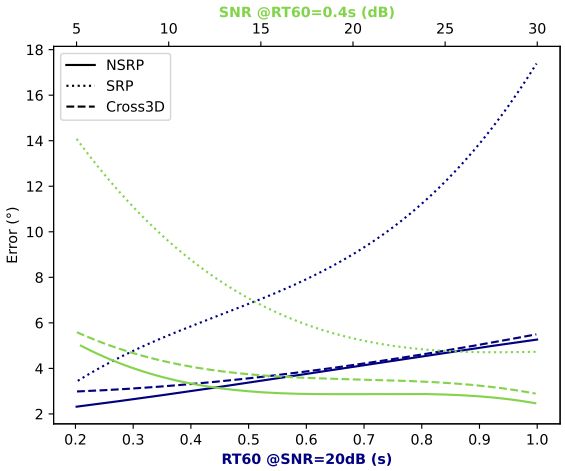

FIGURE 5. Localization error comparison between Neural-SRP, Cross3D and SRP for increasing levels of reverberation and SNR. The curves were smoothed using cubic interpolation.

The method's dependence of localization error to reverberation and SNR is shown in Fig. 5. The error of SRP increases significantly with high reverberation and low SNR, whereas Neural-SRP's error increases less significantly in those conditions. Fig. 5 also shows consistent incremental gains of Neural-SRP in comparison to the Cross3D baseline throughout all reverberation times and SNRs.

As shown in Tab. 4, Neural-SRP was able to be trained on a set of microphone geometries and tested on an unseen microphone geometry with only a small reduction in localization performance. This reduction is however expected, as the prolate spheroid (American football) geometry of the NAO array is not contemplated in the training dataset.

Turning to the multi-source experiment shown in Tab. 5, Neural-SRP achieves an improved localization performance in comparison to the DOANet method, as well as comparable tracking metrics. An important remark is that this increased performance is achieved despite the fact that the DOANet is able to obtain non-causal frame information, as a bidirectional RNN is employed by the latter, which also incurs in a greater number of parameters. A possible explanation for this increased performance is that the Neural-SRP pairwise architecture is more parameter-efficient than the DOANet's global architecture, which has to employ neurons to replicate information for each pair.

Finally, as shown in Tab. 6, the Neural-SRP uses significantly fewer parameters than the other neural baselines, namely, over 6 times fewer parameters than Cross3D and a little over half as many as DOANet. In terms of computational complexity, Neural-SRP is positioned in-between Cross3D and DOANet, being at least 3 times faster than the former, and showing comparable performance with the latter in the case of a 4-microphone array. The proposed method's increase in computational cost is due to its pairwise formulation, which introduces a quadratic dependence with the number of microphone pairs $(M(M-1)/2)$. However, this pairwise formulation also introduces flexibility, as microphone selection procedures such as [47] can be applied to reduce the number of pairs. The pairwise formulation also allows for distributed computing and only requires pairs to be synchronized, which is of particular relevance when using a distributed microphone network [48].

VI. Conclusions

We have presented Neural-SRP, a state-of-the-art localization neural network which is able to overcome limitations of previous neural methods. Besides providing incremental gains in terms of localization performance, Neural-SRP is causal and shows a low computational complexity. Finally, Neural-SRP is the first method that has been shown to work on unseen array geometries.

Future research directions include exploring microphone pair selection methods which may further reduce cost without significantly affecting performance, and extending to locating three or more simultaneous sources.

ACKNOWLEDGMENT

We thank Dr. Matthew Pitkin for helpful discussions.

REFERENCES

[1] S. Li, X. Chang, C. Yang, K. Jiang, Z. Wang, L. Wang, and X. Li, "A Fast Vehicle Horn Sound Location Method with Improved SRP-PHAT," in *2018 IEEE International Conference on Progress in Informatics and Computing (PIC)*, 2018, pp. 435–439.

[2] Y. Li, K. C. Ho, and M. Popescu, "A Microphone Array System for Automatic Fall Detection," *IEEE Trans. on Bio. Eng.*, vol. 59, no. 5, pp. 1291–1301, 2012.

[3] J. Lopez-Morillas, F. J. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, R. Mata-Campos, and V. Montiel-Zafra, "Gunshot detection and localization based on Non-negative Matrix Factorization and SRP-PHAT," in *Sensor Array and Multichan. Sig. Proc. Workshop (SAM)*, 2016, pp. 1–5.

[4] A. Marti, M. Cobos, and J. J. Lopez, "Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2011, pp. 2592–2595.

[5] H. Do and H. F. Silverman, "A robust sound-source separation algorithm for an adverse environment that combines MVDR-PHAT with the CASA framework," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2011, pp. 273–276.

[6] H. Q. H. Dam, H. Ho, and M. H. L. Ngo, "Blind Speech Separation Using SRP-PHAT Localization and Optimal Beamformer in Two-Speaker Environments," *Int. J. of Comp. and Inf. Eng.*, vol. 10, no. 8, pp. 1529–1533, 2016.

[7] C. Wu, L. Zhou, X. Chen, and L. Chen, "Microphone Array Speech Separation Algorithm based on DNN," in *Asia-Pacific Signal and Inform. Process. Assoc. Annual Summit and Conf. (APSIPA)*, 2021, pp. 1305–1310.

[8] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, pp. 276–280, 1986.

[9] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[10] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 2, 1994, pp. 273–276.

[11] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," PhD Thesis, Brown University, 2000.

[12] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A Survey of Sound Source Localization with Deep Learning Methods," *J. Acoust. Soc. Am.*, p. 107–151, 2022.

[13] S. Adavanne, A. Politis, and T. Virtanen, "Differentiable Tracking-Based Training of Deep Learning Sound Source Localizers," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2021, pp. 211–215.

[14] 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.

[15] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2008, pp. 2565–2568.

[16] P. Aarabi, "The Fusion of Distributed Microphone Arrays for Sound Localization," *EURASIP J. on Advances in Signal Process.*, vol. 2003, no. 4, pp. 1–10, 2003.

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

[18] M. B. Çöteli, O. Olgun, and H. Hacıhabiboğlu, "Multiple Sound Source Localization With Steered Response Power Density and Hierarchical Grid Refinement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2215–2229, 2018.

[19] E. Tengan, T. Dietzen, F. Elvander, and T. van Waterschoot, "Multi-source direction-of-arrival estimation using group-sparse fitting of steered response power maps," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2023.

[20] M. Cobos, A. Marti, and J. J. Lopez, "A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, 2011.

[21] L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, M. V. M. Costa, F. M. Gonçalves, A. Said, and B. Lee, "A steered-response power algorithm employing hierarchical search for acoustic

source localization using microphone arrays," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5171–5183, 2014.

[22] J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D Localization and Tracking of Sound Sources Using Beamforming and Particle Filtering," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2006, pp. 4–6.

[23] M. F. Fallon and S. J. Godsill, "Acoustic Source Localization and Tracking of a Time-Varying Number of Speakers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1409–1415, 2012.

[24] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2017, pp. 136–140.

[25] W. He, P. Motlicek, and J.-M. Odobez, "Deep Neural Networks for Multiple Speaker Detection and Localization," in *Proc. Int. Conf. Robotics and Automation*, 2018, pp. 74–79.

[26] L. Perotin, A. Défossez, E. Vincent, R. Serizel, and A. Guérin, "Regression Versus Classification for Neural Network Based Audio Source Localization," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2019, pp. 343–347.

[27] E. Grinstein and P. A. Naylor, "Deep Complex-Valued Convolutional-Recurrent Networks for Single Source DOA Estimation," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[28] E. Grinstein, V. W. Neo, and P. A. Naylor, "Dual input neural networks for positional sound source localization," *EURASIP J. on Audio, Speech, and Music Process.*, no. 32, 2023.

[29] E. Grinstein, M. Brookes, and P. A. Naylor, "Graph neural networks for sound source localization on distributed microphone networks," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2023.

[30] E. Grinstein, T. van Waterschoot, M. Brookes, and P. A. Naylor, "The Neural-SRP method for positional sound source localization," in *Proc. Asilomar Conf. on Signals, Syst. & Comput.*, 2023.

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

[32] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Direction of Arrival Estimation of Sound Sources Using Icosahedral CNNs," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 313–321, 2023.

[33] D. Diaz-Guerra, A. Politis, and T. Virtanen, "Position Tracking of a Varying Number of Sound Sources with Sliding Permutation Invariant Training," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2023, pp. 251–255.

[34] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2017, pp. 2392–2396.

[35] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–10, 2008.

[36] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," in *Proc. Neural Inform. Process. Conf*, 2014.

[37] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2017, pp. 241–245.

[38] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[39] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[40] D. Diaz-Guerra and J. R. Beltran, "Direction of Arrival Estimation with Microphone Arrays Using SRP-PHAT and Neural Networks," in *Sensor Array and Multichan. Sig. Proc. Workshop (SAM)*, 2018, pp. 617–621.

[41] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA Challenge: Acoustic Source Localization and Tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1620–1643, 2020.

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

[43] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2015, pp. 5206–5210.

[44] A. Politis, S. Adavanne, and T. Virtanen, "A Dataset of Reverberant Spatial Sound Scenes with Moving Sources for Sound Event Localization and Detection," in *Detection and Classification of Acoustic Scenes and Events Workshop*, 2020.

[45] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The NIGENS General Sound Events Database," *arXiv*, no. 1902.08314, 2020.

[46] T. Dietzen, E. De Sena, and T. van Waterschoot, "Low-Complexity Steered Response Power Mapping Based on Nyquist-Shannon Sampling," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2021, pp. 206–210.

[47] K. Kumatani, J. McDonough, J. F. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, 2011, pp. 1–6.

[48] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *IEEE Symp. on Comms. and Vehicular Tech. in the Benelux (SCVT)*, 2011, pp. 1–6.