The Neural-SRP method for positional sound source localization

Eric Grinstein*, Toon van Waterschoot†, Mike Brookes* and Patrick A. Naylor*
*Department of Electrical and Electronic Engineering, Imperial College London, U.K.
†Department of Electrical Engineering (ESAT), KU Leuven, Belgium
Email: e.grinstein@imperial.ac.uk

2023 57th Asilomar Conference on Signals, Systems, and Computers | 979-8-3503-2574-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEEECONF59524.2023.10476973

***Abstract*—Steered Response Power (SRP) is a widely used method for the task of sound source localization using microphone arrays, showing satisfactory localization performance on many practical scenarios. However, its performance is diminished under highly reverberant environments. Although Deep Neural Networks (DNNs) have been previously proposed to overcome this limitation, most are trained for a specific number of microphones with fixed spatial coordinates. This restricts their practical application on scenarios frequently observed in wireless acoustic sensor networks, where each application has an ad-hoc microphone topology. We propose Neural-SRP, a DNN which combines the flexibility of SRP with the performance gains of DNNs. We train our network using simulated data and transfer learning, and evaluate our approach on recorded and simulated data. Results verify that Neural-SRP's localization performance significantly outperforms the baselines.**

Index Terms: Sound Source Localization (SSL), Deep Neural Network (DNN), Steered Response Power (SRP), Distributed Microphone Array (DMA)

I. Introduction

This paper focuses on the task of Sound Source Localization (SSL) using signals recorded by a Distributed Microphone Array (DMA). The goal of SSL is to estimate the position of a sound source of interest, such as a human talker, in a room. Applications of SSL include speech enhancement [5], diarization [34], and robot orientation [13]. In turn, a DMA is a network of connected microphones which can be made using custom devices consumer devices such as voice assistants, cell phones, and laptops equipped with one or more microphones. When these devices are wirelessly connected, a DMA is commonly referred to as a Wireless Acoustic Sensor Network (WASN) [4].

A classical approach for SSL is to apply the Steered Response Power (SRP) method, which operates by assigning a likelihood to each location in a grid of candidate source positions. SRP can function on DMAs of any geometry and number of microphones, and has been shown to be robust in moderately reverberant environments [33]. However, its performance is degraded under highly reverberant environments, as its sound propagation model does not model reflections caused by walls and objects within the room.

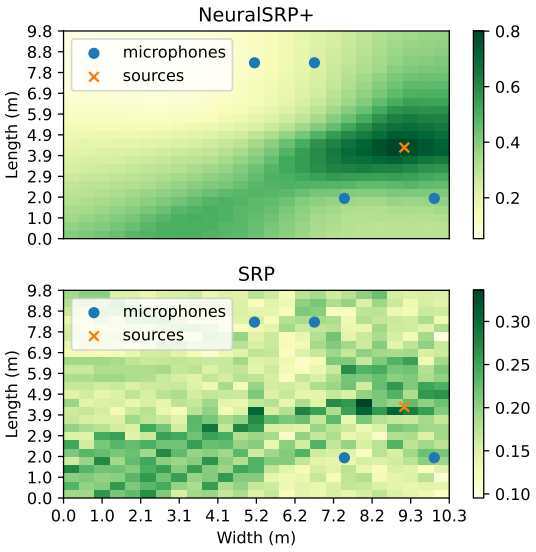


Fig. 1. Neural-SRP and classical SRP output for real recorded signals.

Deep Neural Network (DNN) methods have been proposed to overcome this limitation [19], [7], [1] by using reverberant data for training. However, as most methods were developed for centralized microphone arrays such as a single voice assistant, their application is unsuitable for WASNs where the position and number of microphones may be unknown in advance or dynamically change. A change in number may be caused by the failure of a device, or its position may be moved by a user.

In this work we propose Neural-SRP, a method which combines the advantages of the classical SRP method

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. 956962 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.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 29,2025 at 06:46:06 UTC from IEEE Xplore. Restrictions apply.

and DNNs. As in the classical SRP, this network is able to function on unseen WASN topologies, producing a likelihood grid as its output that is illustrated in the top panel of Fig. 1. As Fig. 1 shows, the maps produced by Neural-SRP are much smoother than classical SRP, resulting in an increased localization performance. This is achieved by training the network on simulated and recorded reverberant data. Another advantage of Neural-SRP is that it does not require calibrated microphone gains, allowing a DMA of heterogeneous devices to be used.

This paper continues as follows. In Sec. II the signal model and scope of this paper is formulated. Relevant background is provided in Sec. III, first by describing the SRP method followed by a description of relevant work in deep learning for SSL. The proposed method is described in Sec. IV, followed by its experimental validation in Sec. V. Finally, results are presented in Sec. VI and Sec. VII concludes this work.

II. Problem statement

Our goal is to estimate the position of a static sound source located at an unknown position $\boldsymbol{p}_s = [p_s^x\, p_s^y\, p_s^z]^T$ within a reverberant room of known dimensions $\boldsymbol{d} = [d^x\, d^y\, d^z]^T$. The source emits a speech signal $s(t)$ at the discrete time index t. Besides the source, M microphones are present in the room, where microphone m has a known position $\boldsymbol{p}_m = [p_m^x\, p_m^y\, p_m^z]^T$, and receives a signal $x_m(t)$ equal to

$$x_m(t) = \sum_{r=0}^{R-1} h_m^r s(t-r) + \epsilon_m(t), \tag{1}$$

where vector $[h_m^0 \dots h_m^{R-1}]^T$ is the Room Impulse Response (RIR) between microphone m and source, and $\epsilon_m(t)$ is measurement noise. In an anechoic room, (1) can be simplified as

$$x_m(t) = a_m s(t-\tau_m) + \epsilon_m(t), \tag{2}$$

where a_m and τ_m respectively represent the attenuation and delay caused by propagation. The anechoic formulation defined in Equation 2 is used by the classical SRP, and does not account for reverberation.

In our method and in the baselines considered in this paper, the microphone signals are sampled and processed in frames of size L, defined as $\boldsymbol{x}_m(t) = [x_m(t-L+1) \dots x_m(t)]^T$ We assume synchronicity between the microphones' digital-to-analog conversion clocks, or that a sampling rate compensation algorithm [8] can be used. Finally, we also define a metadata vector $\boldsymbol{\phi}$ as

$$\boldsymbol{\phi} = [\boldsymbol{p}_1^T \dots \boldsymbol{p}_M^T\, \boldsymbol{d}^T]^T, \tag{3}$$

which specifies the microphone positions and room dimensions, which may be explored as part of the source estimation procedure.

Given the aforementioned data, the goal of our proposed method and baselines is to estimate the 2D coordinates $\hat{\boldsymbol{p}}_s = [\hat{p}_s^x\, \hat{p}_s^y]^T$ of the source, dispensing with the need for estimating its height. A practical scenario for this assumption is work meetings, where all talkers are located at approximately the same height.

III. Related work

A. Steered Response Power

We start by defining the SRP between a candidate source location $\boldsymbol{p} = [p^x\, p^y]^T$ and a pair of microphones (i, j). We will omit the index t for conciseness. The *pairwise SRP* for a candidate location $\boldsymbol{p}$ is defined as [11], [12]

$$\text{SRP}_{ij}(\boldsymbol{p}\,;\, \boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i \star \boldsymbol{x}_j)(\tau_{ij}(\boldsymbol{p})), \tag{4}$$

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

$$\tau_{ij}(\boldsymbol{p}) = (\|\boldsymbol{p}_i - \boldsymbol{p}\| - \|\boldsymbol{p}_j - \boldsymbol{p}\|)/c, \tag{5}$$

that is, the time propagation difference between the microphones located at $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ and the source $\boldsymbol{p}$. c is the speed of sound. In practice, Generalized Cross-Correlation with Phase Transform (GCC-PHAT) [25] is commonly used instead of classical temporal cross-correlation. Finally, the global SRP is defined as the sum between all SRP pairs:

$$\text{SRP}(\boldsymbol{p}\,;\, \boldsymbol{x}) = \sum_{i=1}^{M} \sum_{j=i+1}^{M} \text{SRP}_{ij}(\boldsymbol{p}\,;\, \boldsymbol{x}_i, \boldsymbol{x}_j) \tag{6}$$

This value represents the likelihood of a source being located at a candidate point $\boldsymbol{p}$. In other words, the source location is estimated as $\arg\max_{\boldsymbol{p}} \text{SRP}(\boldsymbol{p})$. In practice, (6) is evaluated for a grid of uniformly spaced points.

B. Neural Networks for SSL

Neural networks have been widely applied for the task of Direction-of-Arrival (DOA) estimation using a centralized microphone array. In contrast with our application, DOA estimation predicts the angles between the source and a reference point in the array, and range is not estimated. In [19], a review of papers is provided. In [28], the choice of output strategy, regression or classification, is investigated. In [7], the phase of the Short Time Fourier Transform (STFT) is used as input feature to a Convolutional Neural Network (CNN), allowing the network to function using uncalibrated microphone gains. In [22], a Gaussian-like output target is introduced, which we adapt for the task of SSL.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 29,2025 at 06:46:06 UTC from IEEE Xplore. Restrictions apply.

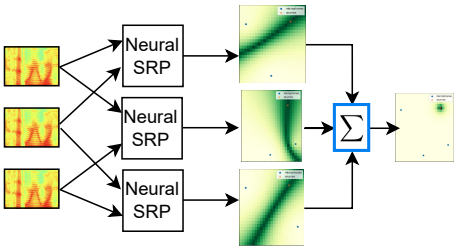


Fig. 2. Example of the Neural-SRP method for three microphones. The magnitude of the STFT instead of its phase is used as input for illustrative purposes.

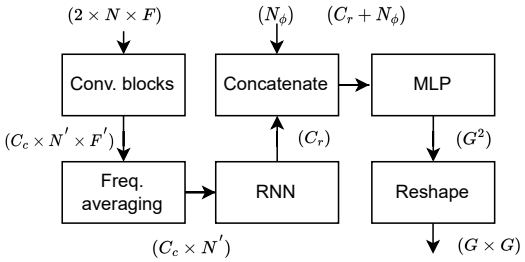


Fig. 3. Architecture of the Neural-SRP network.

Works dedictated to working with a variable number of microphone positions include [16], where an attention-based neural network capable of handling connection failures is proposed for the task of speech enhancement. However, this network is limited to a maximum number of input microphones channels. In [31], a Graph Neural Network (GNN) [24] is also used for speech enhancement. However, this architecure is unsuited for SSL, as they are unable to incorporate the microphone positions in their input.

IV. Neural-SRP

Our proposed method works in a similar way to the classical SRP defined in (6), that is,

$$\text{NSRP}(\boldsymbol{p}\,;\boldsymbol{x}) = \sum_{i=1}^{M}\sum_{j=i+1}^{M} \text{NSRP}_{ij}(\boldsymbol{p}\,;\boldsymbol{x}_i,\boldsymbol{x}_j), \qquad (7)$$

where NSRP represents our neural network architecture used. We note that, in practice, NSRP jointly computes a likelihood map for a square grid $\boldsymbol{P}$ of candidate points of size G^2, as shown in Fig. 3.

An example of our approach's functioning for three microphones is shown in Fig. 2. Note that however, in practice, the candidate positions $\boldsymbol{p}$ are not provided as input, but only used to generate the training targets. The input of the network consists of a pair of signal frames $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and metadata $\boldsymbol{\phi}_{ij} = [\boldsymbol{p}_i^T \boldsymbol{p}_j^T \boldsymbol{d}^T]^T$, that is, a vector concatenating the coordinates of microphones i and j and the room dimensions d.

Fig. 3 shows the building blocks that compose the Neural-SRP architecture, as well as their input and output shapes. Our backbone consists of a Convolutional Recurrent Neural Network (CRNN) [9], which is widely adopted for the task of DOA estimation [6], [1], [17], [18]. A feature of our architecture is its dual-input structure which allows it to incorporate the metadata vector $\boldsymbol{\phi}_{ij}$ and allow it to function on any room and WASN topology.

As in [7], the input of Neural-SRP is the stacked phase of the STFT from the two microphone channels, i.e., a tensor of dimensions $(2, N, F)$, where N and F respectively represent the number of time and frequency bins. A Convolutional Neural Network (CNN) is then applied to these features producing a 3D tensor of C_c channels. The frequency dimension is then averaged, resulting in a 2D matrix which is input to a Recurrent Neural Net (RNN) with Gated Recurring Units (GRUs) [10]. We use the last output stage produced by this network, a feature vector of size C_r and concatenate it to the metadata vector $\boldsymbol{\phi}_{ij}$ of length N_ϕ to produce a *metadata aware* feature. In the field of multimodal fusion [3], the combination procedure of signals and metadata is classified as a *late fusion* approach. Finally, this metadata-aware feature vector serves as input to a Multi-layer Perceptron (MLP), which generates the output likelihood grid of size G^2.

We remark that using a CNN with unitary time kernels and a uni-directional RNN makes our architecture causal, and therefore able to be employed in real-time applications.

A. Training targets

Our network's desired output is a likelihood grid where each candidate grid cell's value is proportional to its distance to the source, that is, $\text{NSRP}(\boldsymbol{p}) \propto |\boldsymbol{p} - \boldsymbol{p}_s|$. We would also like to bound the value in the interval $[0, 1]$, so it can be normalized and interpreted as a probability. In [22], a Pseudo-Gaussian target equal to

$$y_g(\boldsymbol{p}) = e^{-(\|\boldsymbol{p}-\boldsymbol{p}_s\|/\sigma)^2} \qquad (8)$$

is used, which satisfies the aforementioned properties. However, this target is unrealistic for training on a single microphone pair, as source locations share the same cross-correlation function along a hyperbola [21] in anechoic scenarios. We therefore model our target output using a *hyperbolic target grid*, defined for each grid cell as

$$y_h(\boldsymbol{p}) = e^{-(\|\tau_{ij}(\boldsymbol{p})-\tau_{ij}(\boldsymbol{p}_s)\|/\sigma)^2}, \qquad (9)$$

where σ controls the function's decay. $\tau_{ij}(\boldsymbol{p})$ represents the theoretical TDOA between microphones i and j from a source at $\boldsymbol{p}$, as defined in (5). A comparison

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 29,2025 at 06:46:06 UTC from IEEE Xplore. Restrictions apply.

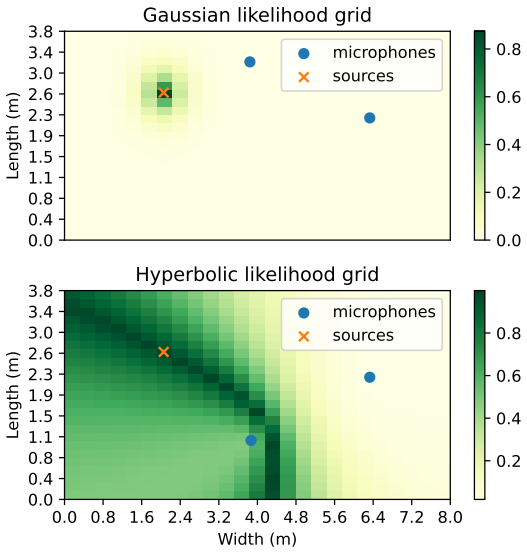


Fig. 4. Example of the hyperbolic grid (below), used for training Neural-SRP, and the alternative Gaussian grid (above).

between both grids can be seen in Fig. 4. Note that (9) is maximized to a value of 1 at the source's location, and decays to 0 as the distance to the source increases. This grid assigns a high likelihood value along a hyperbola branch, hence its name. An interpretation for this target is that the network must learn to estimate a likelihood grid similar to one that would be produced by SRP in an anechoic environment. In other words, the network has to learn to jointly dereverberate and locate the source.

Finally, we define the training loss function as

$$\mathcal{L}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = |\boldsymbol{Y} - \hat{\boldsymbol{Y}}|, \tag{10}$$

the Mean Absolute Error (MAE) between the target grid $\boldsymbol{Y}$ computed using (9) and the network output $\hat{\boldsymbol{Y}}$. The training procedure for Neural-SRP consists of applying a gradient based method search for the network parameters or weights which minimize (10).

B. Anechoic to reverberant transfer learning

We found that training directly on highly reverberant data led to the network getting stuck on local optima values. We were able to overcome this by applying a two-stage training procedure. Firstly, the network is trained using anechoic data. Then, we resume training using a reverberant dataset, which allowed the network to reach substantially lower error values. This procedure is referred to as transfer learning or fine-tuning [30].

V. Experimentation

A. Datasets

This section describes the different datasets used for training and evaluation of Neural-SRP and the baselines. To assess the dependence of Neural-SRP's performance on the number of microphones used, two testing variants of each dataset were produced, one containing 4 microphones and one containing 6 microphones. Furthermore, to train the baseline CRNN, training and validation datasets containing 4 and 6 microphones were used. We emphasize that no combination of microphone and source positions overlap between training, validation and testing datasets. The microphone signals used had a duration of 0.5 seconds.

The **AnechoicSim** dataset is a simulated dataset with no reverberation, used to train the first stage of the network. It consists of 10000 training, 2500 validation and 2500 test samples. The source signal used are speech samples from the VCTK corpus [32]. For each dataset sample, the room's width, length and height are randomly uniformly chosen from the respective intervals of $[3, 10]$, $[3, 10]$ and $[2, 4]$ meters. The source and microphone positions were uniformly sampled within the room's dimensions, with the restriction of each device pair being at least 0.5 m apart.

The **ReverbSim** is a dataset generated using the Image Source method [2] and speech samples from the VCTK corpus, used to train the network at a second stage. Additionally to using the same random room dimension, microphone and source positions ranges of AnechoicSim, the reverberation of each room is set using the *reflectivity-biased* procedure [15], which assigns one absorption coefficient per surface, generating more realistic rooms than creating surfaces with a shared coefficient. It consists of 10000 training, 2500 validation and 2500 test samples.

The **Recorded** dataset [20] contains real recordings from a single room with a high reverberation time of 800 ms containing 40 microphones and 4 loudspeaker positions. The loudspeakers emit sounds from the LibriSpeech [26] corpus. This dataset was used for evaluating the proposed model and baselines, as well as for fine-tuning the neural models. It consists of 250 training samples and 2500 testing samples.

B. Methods and baselines

We evaluate our approach against the classical SRP method using a similar approach to [11]. Furthermore, we experiment training Neural-SRP using two or three stages. The two-stage approach, which we will refer as *NeuralSRP*, consists of training the network using simulated data as defined in Sec. IV-B, firstly by training the network on the AnechoicSim dataset, followed by training it on the ReverbSim dataset. The three-stage approach, denoted *NeuralSRP+*, is further training the *NeuralSRP* model using a small subset of the Recorded

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 29,2025 at 06:46:06 UTC from IEEE Xplore. Restrictions apply.

dataset. Using recorded data to complement synthetic training was used in other works such as [22], as simulated data may not completely match real scenarios.

To compare Neural-SRP against a trained baseline, we propose the CRNN4 and CRNN6 models, which share a similar architecture to and number of parameters to NeuralSRP, but differ in the input and output. CRNN4 and CRNN6 jointly process respectively 4 and 6 microphone signals, and therefore have to be trained specifically for each case. The architecture used is therefore the same as shown in Fig. 3, but the input to the Convolutional blocks is $(4 \times N \times F)$ for CRNN4, and $(6 \times N \times F)$ for CRNN6. Furthermore, a single output grid is produced instead of summing grids from all pairs. For that reason, we train the aforementioned models using the Pseudo-Gaussian target as described in (8).

C. Experiment details

A public repository containing all methods is provided on Github, as well as a demonstration website for data access and reproduction [1]. Pytorch [27] was used as the main deep learning library, along with Pytorch Lightning [14] for abstracting common training routines. Pyroomacoustics [29] was used for generating the AnechoicSim and ReverbSim datasets.

We train the networks for a maximum of 50 epochs with early stopping if the validation loss stops increasing after 3 epochs. We employ a learning rate of 0.0005 using the Adam optimizer [23]. We use a batch size of 16. These parameters were chosen empirically. All grids used are of dimensions 25×25. The neural networks contain four convolutional layers using 1×2 kernels and filter sizes and a MLP containing 3 layers, each of output size 625. The total number of parameters for the neural network methods is around one million. We use a ReLU activation function for all layers except for the output, which uses no activation. The source estimation procedure in the baselines and proposed methods consist of picking the location of the highest value, and GCC-PHAT is used to compute the cross-correlation within the SRP baseline.

VI. Results and discussion

We employ the error $\|\hat{\boldsymbol{p}} - \boldsymbol{p}\|$, i.e. the distance in meters between the predicted and actual source positions, as the main metric of comparison. As a global comparison metric, the average error for all samples is computed. A table comparing our proposed methods and baselines is shown in Tab. I.

The highlighted results in Tab. I show that NeuralSRP and NeuralSRP+ obtain the best results in terms of average localization error, overcoming the classical SRP method. The relative improvement between Neural and classical SRP was of 59, 54, 67, and 34% for the ReverbSim4, Recorded4, ReverbSim6, and Recorded6 datasets respectively. Interestingly, the CRNN baselines were unable to surpass the performance of SRP. An explanation for this performance may be that the Gaussian grid is less effective for training due to its sparse nature. The performance of NeuralSRP increased when using 6 microphones in comparison to 4, as expected.

Dataset	Model	Avg. err. (m)	Std. (m)
ReverbSim4	SRP	1.86	1.46
ReverbSim4	NeuralSRP	**1.17**	0.87
ReverbSim4	CRNN4	2.85	1.53
ReverbSim6	SRP	1.51	1.32
ReverbSim6	NeuralSRP	**0.90**	0.66
ReverbSim6	CRNN6	2.49	1.29
Recorded4	SRP	1.19	1.53
Recorded4	NeuralSRP+	**0.77**	0.66
Recorded4	CRNN4	3.78	1.63
Recorded6	SRP	0.75	1.04
Recorded6	NeuralSRP+	**0.56**	0.46
Recorded6	CRNN6	2.91	1.77

TABLE I
COMPARISON OF THE MEAN AND STANDARD DEVIATION OF THE LOCALIZATION ERROR BETWEEN MULTIPLE DATASETS AND MODELS.

In qualitative terms, an example comparison between grids generated by classical and Neural-SRP can be seen in Fig. 1, showing that while SRP produces a noisy grid with peaks not located at the true source position, Neural-SRP produces a smooth grid with the maximum located at the true source's position.

VII. Conclusion and future work

In this work, we presented Neural-SRP, a method for SSL combining the advantages of the classical SRP method and neural networks. We show that Neural-SRP is able to operate on multiple array geometries, overcoming SRP and a CRNN baseline by a minimum margin of 34% on multiple challenging scenarios. Future work directions include testing Neural-SRP for localizing multiple sources and performing in non-shoebox rooms.

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

References

[1] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pages 1462–1466, 2018.

[2] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950, April 1979.

[3] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379, November 2010.

[4] Alexander Bertrand. Applications and trends in wireless acoustic sensor networks: A signal processing perspective. In *Proc. IEEE Symp. on Comms. and Vehicular Tech. in the Benelux (SCVT)*, pages 1–6, November 2011.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 29,2025 at 06:46:06 UTC from IEEE Xplore. Restrictions apply.

[5] Michael Brandstein and Darren Ward. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2001.

[6] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley. Polyphonic Sound Event Detection and Localization using a Two-Stage Strategy. *Proc. Detect. and Classific. of Acoust. Scenes and Events (DCASE)*, pages 30–34, 2019.

[7] Soumitro Chakrabarty and Emanuel 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)*, pages 136–140, October 2017.

[8] Aleksej Chinaev, Philipp Thüne, and Gerald Enzner. A Double-cross-correlation Processor for Blind Sampling Rate Offset Estimation in Acoustic Sensor Networks. In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pages 641–645, May 2019.

[9] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pages 2392–2396, March 2017.

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *Proc. Neural Inform. Process. Conf*, 2014.

[11] Maximo Cobos, Amparo Marti, and Jose J. Lopez. A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling. *IEEE Signal Processing Letters*, 18(1):71–74, January 2011.

[12] Joseph Hector DiBiase. *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*. PhD thesis, Brown University, 2000.

[13] 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 Trans. Audio, Speech, Language Process.*, 28:1620–1643, 2020.

[14] William Falcon and The PyTorch Lightning team. PyTorch lightning.

[15] Cédric Foy, Antoine Deleforge, and Diego Di Carlo. Mean absorption estimation from room impulse responses using virtually supervised learning. *J. Acoust. Soc. Am.*, 150(2):1286–1299, August 2021.

[16] Nicolas Furnon, Romain Serizel, Slim Essid, and Irina Illina. Attention-based distributed speech enhancement for unconstrained microphone arrays with varying number of nodes. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, June 2021.

[17] Eric Grinstein and Patrick A. Naylor. Deep Complex-Valued Convolutional-Recurrent Networks for Single Source DOA Estimation. In *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, pages 1–5, September 2022.

[18] Pierre-Amaury Grumiaux, Srdan Kitić, Laurent Girin, and Alexandre Guérin. Improved feature extraction for CRNN-based multiple sound source localization. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pages 231–235, August 2021.

[19] Pierre-Amaury Grumiaux, Srdan Kitić, Laurent Girin, and Alexandre Guérin. A Survey of Sound Source Localization with Deep Learning Methods. *J. Acoust. Soc. Am.*, 152(1):pp107–151, September 2021.

[20] Shanzheng Guan, Shupei Liu, Junqi Chen, Wenbo Zhu, Shengqiang Li, Xu Tan, Ziye Yang, Menglong Xu, Yijiang Chen, Jianyu Wang, and Xiao-Lei Zhang. Libri-adhoc40: A dataset collected from synchronized ad-hoc microphone arrays. In *Proc. Asia-Pacific Signal and Inform. Process. Assoc. Annual Summit and Conf. (APSIPA)*, 2021.

[21] F. Gustafsson and F. Gunnarsson. Positioning using time-difference of arrival measurements. In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2003.

[22] Weipeng He, Petr Motlicek, and Jean-Marc Odobez. Deep Neural Networks for Multiple Speaker Detection and Localization. In *Proc. Int. Conf. Robotics and Automation*, pages 74–79, May 2018.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *arXiv: 1412.6980*, January 2017.

[24] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proc. Int. Conf. on Learning Representations*, February 2017.

[25] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust., Speech, Signal Process.*, 24(4):320–327, August 1976.

[26] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pages 5206–5210, April 2015.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Proc. Neural Inform. Process. Conf*, volume 32. Curran Associates, Inc., 2019.

[28] Lauréline Perotin, Alexandre Défossez, Emmanuel Vincent, Romain Serizel, and Alexandre 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)*, pages 343–347, October 2019.

[29] 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)*, pages 351–355, 2018.

[30] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A Survey on Deep Transfer Learning. In *Proc. Int. Conf. on Artificial Neural Networks and Machine Learning (ICANN)*, pages 270–279, Cham, 2018. Springer International Publishing.

[31] Panagiotis Tzirakis, Anurag Kumar, and Jacob Donley. Multi-Channel Speech Enhancement using Graph Neural Networks. *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, February 2021.

[32] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), November 2019.

[33] Cha Zhang, Dinei Florencio, and Zhengyou Zhang. Why does PHAT work well in lownoise, reverberative environments? In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pages 2565–2568, March 2008.

[34] Siqi Zheng, Weilong Huang, Xianliang Wang, Hongbin Suo, Jinwei Feng, and Zhijie Yan. A Real-Time Speaker Diarization System Based on Spatial Spectrum. In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pages 7208–7212, June 2021.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on July 29,2025 at 06:46:06 UTC from IEEE Xplore. Restrictions apply.