International Journal of Engineering Sciences & Emerging Technologies, Oct. 2023.
ISSN: 22316604 Volume 11, Issue 2, pp: 24-33 ©IJESET

A Systematic Review of Neural Networks Based Sound Source Localization

Himani Bhayana, Ritu Boora, Manisha Jangra

Department of Electronics and Communication Engineering,
Guru Jambheshwar University, Hisar, India
himanibhayana.198@gmail.com
rituboora@gmail.com
Manisha.05.86@gmail.com

Abstract
Neural Network based sound source localization has rapidly evolved over the last few years. However, there is a continuous rise in demand for more accurate, reliable, and computationally efficient methods that can be effectively implemented in harsh acoustic conditions. This article aims to collectively examine the existing source localization techniques implemented with neural networks. It provides discussions on the architectures, learning strategy, features extracted, number of sources, dataset, and the performance of these methods. A table summarizing the literature review is supplied at the end to enable a quick search of approaches with a specific set of goal features.

Keywords: *CNN, RNN, LSTM, Localization, Neural Network, Sound Source*

1. Introduction

Sound source localization (SSL) is the task of approximating the position of one or more sound sources with respect to random reference positions. Typically, this reference point is the location of a recording source array. Despite the fact that researchers have made great strides in the field of source localization [1–3] it still remains an emerging field due to its rising demand in modern SSL-based applications such as human-machine interaction, hearing aids, automated surveillance, automated camera steering in videoconferencing, teleconferencing, etc. In these applications, SSL can be used for source separation [4], speech enhancement [5], speech recognition [6], control robot movement, source tracking, and many more. These applications demand efficient and robust localization techniques with low computational burden in complex environments[7].

The existing SSL techniques can be classified based on any one of these four parameters: design principle of techniques, sensor array used, localization in 3D space, and the number of active sources. Based upon their design principle, these SSL techniques can be broadly categorized into two groups namely conventional techniques and Neural Network based techniques. The conventional methods are basically based on either of these: Beamforming[8], Time Difference of Arrival [9] or Steered Response Power [10, 11]. However, in the last few years, neural network-based techniques have seen surge in their applications. Hence, this work aims to present the survey of the Neural Network based SSL techniques.

Several authors have presented the application of Neural Networks (NN) such as Artificial Neural Networks (ANN)[12], Convolutional Neural Networks (CNN)[13], Recurrent Neural Networks (RNN)[[14]], Long Short Term Memory (LSTM) and their combinations for source localization. These localization task can be formulated as regression or classification problem [15] .In the case of the

DOI: 10.5281/zenodo.10434167

regression problem, it aims to predict the continuous values whereas the discrete values solutions are found in case of classification. The second categorization of SSL methods is based on number of sensors used and their placement. They can be classified into monoaural, binaural, tri-aural, tetra-aural, and multi-aural microphone arrays based on the number of microphones used. These sensors can be placed in several geometries such as linear, circular, non-coplanar, tetrahedral, and so on to achieve optimum performance[16]. Among these techniques, binaural can rarely be replaced exclusively in the fields associated with human hearing such as hearing aids, humanoid robots, and so on. These techniques rely on the data input from solely two microphones for localization which makes them efficient in terms of hardware required. They usually extract one or more of auditory cues namely Interaural Level Difference (ILD), Interaural Time Difference (ITD), and Interaural Phase Difference (IPD) of the received signals as use it as input to NN techniques[17–19].

The third classification of SSL techniques can be made on the number of space dimensions in which source is localized. In 1D localization, it is typically the azimuth angle of the source which is estimated from the given signals [20]. Azimuth and elevation are the predicted dimensions in 2D [21]whereas in 3D, the polar or Cartesian coordinates of the source are estimated [22]. Several SSL-based applications need direction of arrival (DOA) of the source which requires location estimation in only 1D or 2D only such as videoconferencing, speech enhancement etc. The other applications may need the cartesian coordinates of the source such as surveillance, robot movement control etc. The fourth essential SSL parameter is the number of inputs in a recorded mixing signal. In several work, the number of active sources are not known prior to implementation of SSL methodology, in that case source is detected before localization [23, 24].

This paper presents a survey of Neural Network based SSL techniques. It aims to organize the literature work on SSL and present it is a convenient form for the new researchers in this field.

This script is further organized as follows: The section 2 describes the literature-recommended neural network architectures for solving the SSL challenge. It includes the networks' layering scheme, with a progressive network and complicated strategy. The first subsection is based on learning techniques which are used for datasets. The second section is based on Neural Network techniques for SSL. It is followed by DNN, CNN, RNN's and their hybrid models in consecutive subsections. Finally, the study of existing techniques are presented in table form.

This paper presents a systematic survey of the SSL literature using different architectures. We categorize and discuss the various approaches of the employed architectures and addressed number of sources for localization system. In other words, we provide a taxonomy of the recently released ML-based SSL literature. At the end of the paper, we present a summary of this review in the form of detailed table including all features used with model and evaluated results.

2. ML BASED SSL TECHNIQUES

2.1 Implementation

The source localization is designed in two stages to detect the direction of an oncoming sound wave. The first step is to generate datasets, and the second is to train with network for localization estimation. This article discusses many strategies for sound source localization in which key feature is Cue extraction (ILD,IPD,ITD, and GCC-PHAT etc. In contrast to humans, which use binaural sound localization techniques, robotic ears first used numerous microphones and different array configurations for SSL, employing both conventional localization algorithms which contains number of hidden layers and provide a calculative approach for localization angle , Accuracy and Error.

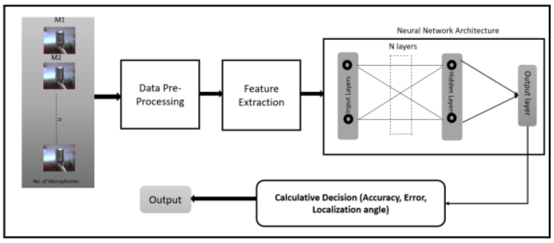

Fig 1. Structural implementation for localization.

2.2 Learning Techniques

In general, selecting a training methodology for a neural network to carry out a specific task depends on the kind and quantity of data that is available where learning techniques are used. They are useful when complicated functions are used to describe the relationship between the signals from the microphone array and the characteristics retrieved from those signals. In the neural network-based SSL literature, most of the systems rely on supervised, semi-supervised, or weakly supervised learning. Supervised (S) learning has become a popular technique to learn representations from massive unlabelled datasets. The term "self-supervision" refers to techniques that develop representations of the data using "proxy learning tasks," where the learning process is influenced by patterns in the data. Here, the main objective is to develop mappings between input examples and low-dimensional representations [25]. The drawback of supervised training is that it uses a lot of labelled training data, whereas only a small number of real-world datasets can be gathered for SSL. These datasets are insufficient for reliable Deep Learning (DL) model training. Semi-supervised (SS) learning is when some of the learning is done under supervision and some of it is done unsupervised. The network is often trained using labelled data before being improved (or fine-tuned) via unsupervised learning, which is, without labels [26]. Work in Refs. [27, 28] employed a weakly supervised (WS) training method. The authors improved a pre-trained neural network in this study by changing the cost function, which is the number of known sources to account for weak labels. There are numerous approaches available in addition to the above succinctly mentioned learning methods for determining the DOA estimations.

2.2.1. Neural-based SSL techniques

Deep Neural Network (DNN): Deep Neural Networks are strong models that can automatically recognize and take advantage of the relationship between the location of sources and the availability of information. A sufficient number of representative training examples are used to train the models. Since the feature extractor module can be related to traditional processing, data-driven DL approaches have the potential to replace traditional methods based on a signal procedure and model entirely, or at least in part. This makes SSL-based applications more attractive[29]. Deep learning models effectively adapt to the previously presented training data without explicitly imposing any such assumptions. Deep neural networks and spectral source models are combined in Ma et al.'s[30] model of binaural localization. The model conducts azimuth estimation by combining the target source and noise with DNN, where ITD, CCF, and ILD feature vectors are used. The model has a less than 5% error rate and is resistant to noise and reverberation settings. Using DNN, He et al.[31] carried out multiple speaker detection and localization, which uses likelihood-based coding and the Generalized Cross Correlation Phase Transform (GCC-PHAT) as input features. With 90% accuracy, the model calculated 3D sound-based localization. Binaural localization based on DNN and preferred propagation clustering in incompatible HRTF settings is proposed by Wang et al. [32] to enhance the DNN model's capacity for generalization. It estimates azimuth with 63% accuracy. Using cues like ILD and CCF, work in Ref [33]suggests fusing CNN and DNN. The output layer was applied after the concatenation of the front-back classification performed by CNN and the azimuth estimation performed by DNN. With 83% accuracy, the model performed well in reverberant and noisy situations.

Convolution Neural Network (CNN): CNN is the most popular source localization and classification approach, often known as a convNet [34]. It is one of the key developments in the field of machine

DOI: 10.5281/zenodo.10434167

leaning that helped to pave the way for the DNN renaissance. It consists of pairs of convolution and max pooling layers. The convolution layer applies a series of filters that operate on discrete local regions of the input where they are repeated over the entire input space. According to work in Ref. [25], utilizing a CNN increased overall Direction of Arrival (DoA) classification accuracy by two times when compared to a conventional method called Steered Response Power with Phase Transform (SRP-PHAT) in low signal-to-noise ratio settings. This work is further extended to localize multiple speakers in work [35] where the DOA estimate approach has a reasonably good degree of adaptability to unknown acoustic situations. This method is highly dependent on the time-varying source signal. However, the model is trained using phase component of the input signal's Short-Time Fourier Transform (STFT) which slowed down the training of the network. Li et al.[36] worked around this problem by combining CNN with Long Short-Time Memory (LSTM) which further improved the localization accuracy by 80-90%. Semi-supervised localization using deep generative modelling and Variational Autoencoders (VAE) is proposed by Bianco et al. [37]. Using labelled and unlabelled HRTF samples, VAE provides the relative transfer function phase that the DOA classifier compares against. In comparison to CNN, VAE-based localization is more accurate and uses SRP-PHAT systems. It conducts an azimuth estimation from -90 to 90º with a 2.9 to 4.2% error rate. Based on a weekly supervised learning approach that can be modelled using a small number of labelled samples and a larger collection of unlabelled samples, Opochinsky et al. [38]performs binaural localization. It estimates azimuth from 0 to 180 degrees with an error of under 10%. CNN network for binaural localization has been implemented in Ref. [39]based on ITD using Grouped Cross Correlation Function (GCCF) and Encoded Cross Correlation Function (ECCF). It developed two models for TDOA estimation: GCC net grouped and GCC net encoded, and trained them in three different environments: anechoic rooms, multi-conditional training, and realistic settings. It estimates azimuth between -80 and 80 with a 0.1% to 0.25% error rate.

Convolutional Recurrent Neural Network (CRNN) - RNNs are neural networks made specifically for modelling temporal data sequences [14]. RNNs theoretically have the ability to simulate long-term temporal dependencies, but due to the large number of time steps required, the gradients disappear before they reach the first time steps during optimization. CNN has been combined with RNN which is often used to model sequential data such as source localization. In this, the last convolutional layers of a CNN are changed to create a CRNN, which is known as a modified CNN. In comparison to past architectures that only used CNN, work in [40]showed CRNN, which also uses a GRU (gate recurrent unit), performs better in terms of training time and parameter count. Convolutional Recurrent Neural Network based four-microphone array architecture is suggested by Grondin et al.[19] for azimuth and elevation localization. Using CRNN, work in ref. [22]suggested event localization model which does not rely on configuration of the array because it employs the phase and magnitude components of each channel separately. It can be applied to any type of microphone array design, is resistant to unidentified DOA, and is capable of detecting multiple DOA.

Long Short Term Memory (LSTM): The LSTM was proposed to resolve the problem of vanishing gradients of RNN methods[41].The network contains three layers: a hidden layer with four LSTM blocks, a visible layer with one input and an output layer that predicts value. The recurrent hidden layer of the LSTM has unique components known as memory blocks. These memory blocks also comprise memory cells with self-connections that store the network's temporal state[42].The original architecture includes input gates and output gates for every memory block that regulate the flow of information. Ninad et al.[43] worked on ILD, and the Mel Frequency Cepstral Coefficients (MFCC) are used to train the LSTM-RNN network when the signal from both microphones is detected. The network next learns to determine the direction of the sound signal and does azimuth estimation by extracting distinguishing characteristics from the MFCC. In this work, testing accuracy for 10 and 450 precisions in azimuth has been found to be 82% and 95%, respectively.

DOI: 10.5281/zenodo.10434167

Table 1. A summary of existing NN based technique

Author	**Comparative study of different Architecture with Error, Accuracy and Localization angle.**								
	Architecture	**Dataset**	**Nos**	**LT**	**Output Types**	**Features used**	**ER**	**AC**	**Localization angle (Az: Azimuth, El: Elevation)**
[44]	NN	Roomsim using ISM, Recorded dataset	1	S	R	Frequency-dependent ITD& ILD; cue filtering to reduce reverberation	Sim:20º - 50º, Recorded50º-60 º	-	Az:(-45º: -45º)
[37]	Deep Generative modeling with -VAE based on CNN	DTU and MIR dataset for IR & Libri Speech corpus	1	SS	C	RTF-Phase Sequence	(MSE) 7.81º-3.00º(DTU) And 12.9º-3.11º(MIR)	52-80.5%(DTU) 69.3-84.4%(MIR)	Az: (-90º:90º)
[32]	DNN	CIPIC, RIEC for HRTF	1	S	R	Clustering HRTF, ILD and CCF	2.6º	60.43% (CIPIC), 55% (RIEC)	Az: (-80º:80º)
[30]	DNN –Full head movement	Surrey BRIR & BRIR TU Berlin	3	S	C	ILD and CCF	0.25%	96%	Az: (0:360º)
[45]	CAR-FAC cochlear system: Cascade of Asymmetric Resonators with Fast-Acting Compression Deep- CNN	Recorded dataset, austalk	1	S	R	ITD, IPD and spectral cues	RMSE =3.680º	-	Az: (0-180º)
[38]	DNN based on stochastic combination of triplet-ranking loss for the unlabeled samples and physical loss for the anchor samples,	Simulated datasets using ISM	1	WS	R	Relative Transfer Function (RTF)	30º - 40º	-	Az: (0-180º)
[39]	CNN	CIPIC for HRTF and NOISEX-92	1	S	R	CCF-grouped & CCF-encoded.	TDOA error (0.143ms sim RT 60=0.8ms) 0.279ms for realistic env.	GCC encoded performs better	-
[46]	Fusion of CNN (for front back classification) and DNN	TIMID, AIR, NOISEX-92	1	S	R	ILD and CCF	-	83.43%	Az: (0-360º)

DOI: 10.5281/zenodo.10434167

[47]	TF-CNN with Multitask learning	CIPIC - HRTF database, TIMID	1	S	R	HRTF, IPD and ILD	-	85- 90%	Az; (-80°:80°) El: (-45° to 90°)
[48]	CRNN+SA	DCASE 2020 (FOA & MIC)	1	S	R	Log-Mel magnitude spectrogram, IPD from acoustic intensity (FOA) vector and GCC-PHAT (MIC)	190 (FOA)& 18.20 (MIC)	-	-
[43]	LSTM-RNN	Real dataset at 10 and 450	1	S	R	Mel Frequency Cepstral Coefficients (MFCC) of ILD	-	At 10: 82% & at 450: 95%	-
[22]	CRNN	Synthetic dataset	3	S	R &C	Phases and magnitudes of spectrograms	-	91%	Az:(90°:90°), El: (-60°:-60°)
[49]	CNN	Simulated with ISM, HINT, TIMID, recorded data with smartphone	2	S	C	STFT (real + img)	<20%	83-89%	Az: (0:180°)
[50]	Auto encoder - decoder with explicit transformation layer	Synthetic and simulated	3	SS	C	Real +imaginary spectrograms	Rmse-0.11, F1 score is 0.80	-	-
[51]	CRNN+RA +multi-scale densely connection (MDC)	TAU-NIGENS, ANSIM, REAL, MANSIM, SSEA, MREAL	1	S	R	Log-Mel spectrogram +sound intensity vector	0.535±0.025°	Fr.90.2, DE 14.8	-
[52]	SELD-TCN	TUT, ANSYN, MANSYN, REAL, and MREAL	1	S	R	Phase +Magnitude spectrogram	0.67	-	-

For effective comparison of the performance of state-of-the art methods, these methods should have been tested on centralized environment and similar dataset However, as shown in the table, these methods have been tested on different dataset. Several Authors have taken different speech dataset, Noise dataset and impulse response datasets while others have recorded dataset. Due to lack of availability of sufficient labelled data for training, several authors have even preferred training the model with simulated data. This paper presents the comparison of these methods broadly in terms of architecture, features and their accuracy.

Here, it is noted that the among the two output strategies namely classification and regression, the latter is the commonly preferred by researchers. This is mainly due the fact that SSL is mainly considered as a

DOI: 10.5281/zenodo.10434167

regression problem. Further, we can see that most of the work is limited to azimuth angle localization and moreover only in the front of the microphones. A very few authors have considered the front-back ambiguity as a challenge. Author in work [46]has suggested front- back localization as a classification problem and then localized the source as a regression task while another has considered full head movement for 3600 localizations. From the table it may be depicted that, the existing work is limited to localizing the single source. However, in practical environment there are more than one source active at a time. However, in real world, there are more than one active source at a given time and thus open the doors for detection and localization of multiple active sources simultaneously in 3D. The table shows that most of the researchers have opted for supervised learning and hence these localization systems may fail to accurately localize when there are frequent changes in the environment. This leaves a scope for the unsupervised or semi-supervised learning approaches to take over the SSL field.

The popularly used architecture are based on DNN and CNN so far which may be due the fact that they have ability to recognize all potential interactions between predictor variables and to implicitly detect complicated nonlinear relationships between dependent and independent variables with availability of different training. However, it is also noted that LSTM-RNN is the one of the growing architectures that has shown improvement in accuracy compared to other methods with less error rate. The commonly used feature that are used for training the model are ILD, ITD and CCF. The authors have preferred the combination of two or more features to improve their performance.

The estimate accuracy for the LSTM-RNN with MFCC of ILD is found to be 95% whereas for other architecture, it fluctuates 80-90%. The DNN architecture with full head movement has shown the highest accuracy of 96% while using ILD and CCF.

3. Conclusion

In this study, we have extensively reviewed the literature on SSL techniques based on neural network and localization approaches developed in the Robotics field during the last years. Various methodologies were presented, and their applicability to robotics was examined. The application of the high-resolution CNN approach to broadband signals necessitates specific consideration due to limited processing resources and the existence of noisy environment. LSTM based architecture have gained the popularity in the last few years. These models are preferably trained with combination of ITD, ILD and CCF. However, most of these techniques have been implemented on a single source in 1D localization. Moreover, most of these techniques are based on supervised learning that demands large, labelled data. Hence, there is a need to make these techniques semi-supervised and extent them for 3D localization. Accuracy and RMSE needs to be improved in scenarios when multiple sources are active at a time. All these upcoming developments will lead to lively debates among the Signal Processing, scientific communities of Acoustics Robotics, but also Psychoacoustics and Physiology. Then, we hope that this assessment of accessible approaches to the "low-level" stage of localizations will inspire new members to join the thriving field of Robot Audition.

References

[1] S. Argentieri, P. Danès, and P. Souères, “A survey on sound source localization in robotics: From binaural to array processing methods,” Comput. Speech Lang., vol. 34, no. 1, pp. 87–112, Nov. 2015, doi: 10.1016/J.CSL.2015.03.003.

[2] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, “A survey of sound source localization methods in wireless acoustic sensor networks,” Wirel. Commun. Mob. Comput., vol. 2017, p. 24, 2017, doi: 10.1155/2017/3956282.

[3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust Localization in Reverberant Rooms,” in Microphone Arrays. Digital Signal Processing, Springer, Berlin, Heidelberg, 2001, pp. 157–180. doi: 10.1007/978-3-662-04619-7_8.

[4] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, “Multi-Microphone Speaker Separation based on Deep DOA Estimation,” in 2019 27th European Signal Processing Conference (EUSIPCO), 2019. doi: https://doi.org/10.23919/EUSIPCO.2019.8903121.

DOI: 10.5281/zenodo.10434167

[5] A. Xenaki, J. Bünsow Boldt, and M. Græsbøll Christensen, "Sound source localization and speech enhancement with sparse Bayesian learning beamforming," J. Acoust. Soc. Am., vol. 143, no. 6, pp. 3912–3921, 2018, doi: 10.1121/1.5042222.

[6] H. Y. Lee, J. W. Cho, M. Kim, and H. M. Park, "DNN-Based Feature Enhancement Using DOA-Constrained ICA for Robust Speech Recognition," IEEE Signal Process. Lett., vol. 23, no. 8, pp. 1091–1095, 2016, doi: 10.1109/LSP.2016.2583658.

[7] C. Evers et al., "The LOCATA Challenge: Acoustic Source Localization and Tracking," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 28, pp. 1620–1643, 2020, doi: 10.1109/TASLP.2020.2990485.

[8] A. Kujawski, G. Herold, And, and E. Sarradj, "A deep learning method for grid-free localization and quantification of sound sources," J. Acoust. Soc. Am. 146, EL225, vol. 225, no. 2019, 2021, doi: 10.1121/1.5126020.

[9] R. Boora and S. K. Dhull, "A TDOA-based multiple source localization using delay density maps," Sādhanā, vol. 45, no. 204, pp. 1–12, 2020, doi: 10.1007/s12046-020-01453-8.

[10] R. Boora and S. K. Dhull, "Iterative Modified SRP-PHAT with Adaptive Search Space for Acoustic Source Localization Iterative Modified SRP-PHAT with Adaptive Search Space for Acoustic Source," IETE Tech. Rev., vol. 39, no. 1, pp. 28–36, 2022, doi: 10.1080/02564602.2020.1819895.

[11] R. Boora and S. K. Dhull, "Performance Evaluation of Iterative SRP-PHAT Techniques for Acoustic Source Localization," in Proceedings of First International Conference on Computational Electronics for Wireless Communications, 2022, vol. 329, pp. 403–418.

[12] P. Castellini, N. Giulietti, N. Falcionelli, A. F. Dragoni, and P. Chiariotti, "A neural network based microphone array approach to grid-less noise source localization," Appl. Acoust., vol. 177, p. 107947, 2021, doi: 10.1016/j.apacoust.2021.107947.

[13] Z. Wang, N. Li, T. Wu, H. Zhang, and T. Feng, "Simulation of Human Ear Recognition Sound Direction Based on Convolutional Neural Network," J. Intell. Syst., vol. 30, no. 1, pp. 209–223, 2020, doi: 10.1515/jisys-2019-0250.

[14] I. G. Y. Bengio and A. Courville, Deep Learning. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[15] L. Perotin, A. Defossez, E. Vincent, R. Serizel, and A. Guerin, "Regression versus classification for neural network based audio source localization," IEEE Work. Appl. Signal Process. to Audio Acoust., vol. 2019-Octob, pp. 343–347, 2019, doi: 10.1109/WASPAA.2019.8937277.

[16] D. Desai and N. Mehendale, "Sound Source Localization Systems : A review," Arch. Comput. Methods Eng., vol. 29, no. 7, pp. 4631–4642, 2022, doi: 10.1007/s11831-022-09747-2.

[17] C. Rascon and I. Meza, "Localization of sound sources in robotics : A review," Rob. Auton. Syst., vol. 96, pp. 184–210, 2017, doi: 10.1016/j.robot.2017.07.011.

[18] A. M. D. Pavlidi, A. Griffin, M. Puigt, "Real-Time Multiple Sound Source Localization and Counting using a Circular Microphone Array," IEEE Trans. Audio, Speech, Lang. Process., vol. 21, no. 10, pp. 2193–2206, 2013, doi: 10.1109/TASL.2013.2272524.

[19] F. Grondin and F. Michaud, "Lightweight and Optimized Sound Source Localization and Tracking Methods for Open and Closed Microphone Array Configurations," Rob. Auton. Syst., vol. 113, 2018, doi: http://dx.doi.org/10.1016/j.robot.2019.01.002.

[20] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, "On sound source localization of speech signals using deep neural networks," in Advances in Acoustics - DAGA 2015: 41st Annual Conference on Acoustics, March 16-19, 2015 in Nuremberg. Berlin: German Society for Acoustics e.V, 2015, pp. 1510–1513. doi: https://doi.org/10.14279/depositonce-8779.

[21] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector," in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), 2018, pp. 241–245. doi: 10.1109/IWAENC.2018.8521403.

[22] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in European Signal Processing Conference, 2018, vol. 2018-Sept, pp. 1462–1466. doi: 10.23919/EUSIPCO.2018.8553182.

[23] C. R. Landschoot and N. Xiang, "Model-based Bayesian direction of arrival analysis for sound sources using a spherical microphone array," J. Acoust. Soc. Am. 146, vol. 146, pp. 4936–4946, 2019, doi: 10.1121/1.5138126.

[24] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," IEEE Trans. Signal Process., vol. 58, no. 1, pp. 121–133, 2010, doi: 10.1109/TSP.2009.2030854.

DOI: 10.5281/zenodo.10434167

[25] S. Chakrabarty and A. P. Habets, “Broadband doa estimation using convolutional neural networks trained with noise signals,” in 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 2017, pp. 136–140. doi: 10.1109/WASPAA.2017.8170010.

[26] R. Takeda, Y. Kudo, K. Takashima, Y. Kitamura, and K. Komatani, “Unsupervised Adaptation of Neural Networks for Discriminative Sound Source Localization with Eliminative Constraint,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 3514–3518. doi: 10.1109/ICASSP.2018.8461723.

[27] Z. He, D. Liu, H. He, D. Barber, and J. Li, “Tracking by Animation : Unsupervised Learning of Multi-Object Attentive Trackers,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1318–1327. doi: 10.1109/CVPR.2019.00141.

[28] W. He, P. Motlicek, and J. M. Odobez, “Adaptation of Multiple Sound Source Localization Neural Networks with Weak Supervision and Domain-adversarial Training,” in 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 770–774. doi: 10.1109/ICASSP.2019.8682655.

[29] T. G. Dvorkind and S. Gannot, “Time difference of arrival estimation of speech source in a noisy and reverberant environment,” Signal Processing, vol. 85, pp. 177–204, 2005, doi: 10.1016/j.sigpro.2004.09.014.

[30] N. Ma, T. May, and G. J. Brown, “Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localization of Multiple Sources in Reverberant Environments,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 25, no. 12, pp. 2444–2453, 2017, doi: 10.1109/TASLP.2017.2750760.

[31] W. He, P. Motlicek, and J. M. Odobez, “Deep Neural Networks for Multiple Speaker Detection and Localization,” in Proceedings - IEEE International Conference on Robotics and Automation, 2018, pp. 74–79. doi: 10.1109/ICRA.2018.8461267.

[32] J. Wang, J. Wang, K. Qian, X. Xie, and J. Kuang, “Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition,” . EURASIP J. Audio, Speech, andMusic Process., vol. 4, 2020, doi: https://doi.org/10.1186/s13636-020-0171-y.

[33] S. Jiang, L. Wu, P. Yuan, Y. Sun, and H. Liu, “Deep and CNN fusion method for binaural sound source localisation,” vol. 2020, no. Acait 2019, pp. 511–516, 2020, doi: 10.1049/joe.2019.1207.

[34] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–44, 2015, doi: 10.1038/nature14539.

[35] S. Chakrabarty, S. Member, A. P. Habets, and S. Member, “Multi-Speaker DOA Estimation Using Deep Convolutional Networks Trained With Noise Signals,” vol. 13, no. 1, pp. 8–21, 2019, doi: 10.1109/JSTSP.2019.2901664.

[36] X. Z. ∗ and H. L. Qinglong Li, “Online Direction of Arrival Estimation Based on Deep Learning,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 2616–2620. doi: 10.1109/ICASSP.2018.8461386.

[37] M. J. Bianco and S. Gannot, “Semi-Supervised Source Localization in Reverberant Environments With Deep Generative Modeling,” in IEEE Access, 2021, vol. 9, pp. 84956–84970. doi: 10.1109/ACCESS.2021.3087697.

[38] G. C. Renana Opochinsky, Bracha Laufer-Goldshtein, Sharon Gannot, “Deep Ranking-Based Sound Source Localization,” in 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2019, pp. 283–287. doi: 10.1109/WASPAA.2019.8937159.

[39] H. Liu, P. Yuan, B. Yang, and L. Wu, “Robust interaural time difference estimation based on convolutional neural network,” in IEEE International Conference on Robotics and Biomimetics, ROBIO 2019, 2019, pp. 352–357. doi: 10.1109/ROBIO49542.2019.8961817.

[40] K. Choi, G. Fazekas, M. Sandler, and K. Cho, “Convolutional recurrent neural networks for music classification,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2017, pp. 2392–2396. doi: 10.1109/ICASSP.2017.7952585.

[41] S. Hochreiter, “Long Short-Term Memory,” Neural Comput., vol. 9, pp. 1735–1780, 1997, doi: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[42] H. Sak, A. Senior, and F. Beaufays, “Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition.,” in Arxiv, 2014, pp. 1–5.

[43] D. Desai and N. Mehendale, “Robotic Ear : Audio Signal Processing for Detecting Direction of Sound,” in SSRN Electronic Journal, 2021, p. 14. doi: 10.2139/ssrn.3891347.

[44] K. Youssef, S. Argentieri, and J. L. Zarader, “A learning-based approach to robust binaural sound localization,” in IIEEE/RSJ International Conference on Intelligent Robots and Systems, 2013, pp. 2927–2932. doi: 10.1109/IROS.2013.6696771.

DOI: 10.5281/zenodo.10434167

[45] Y. Xu, S. Afshar, R. K. Singh, R. Wang, A. Van Schaik, and T. J. Hamilton, “A binaural sound localization system using deep convolutional neural networks,” Proc. - IEEE Int. Symp. Circuits Syst., vol. 2019-May, pp. 2–6, 2019, doi: 10.1109/ISCAS.2019.8702345.

[46] C. Pang, H. Liu, and X. Li, “Multitask learning of time-frequency CNN for sound source localization,” IEEE Access, vol. 7, pp. 40725–40737, 2019, doi: 10.1109/ACCESS.2019.2905617.

[47] H. Phan, L. Pham, P. Koch, N. Q. K. Duong, I. Mcloughlin, and A. Mertins, “Audio Event Detection and Localisation with Multitask Regression Network.,” 2020.

[48] A. Küçük, A. Ganguly, Y. Hao, and I. M. S. Panahi, “Real-Time Convolutional Neural Network-Based Speech Source Localization on Smartphone,” IEEE Access, vol. 7, pp. 169969–169978, 2019, doi: 10.1109/ACCESS.2019.2955049.

[49] G. Le Moing, P. Vinayavekhin, D. J. Agravante, T. Inoue, J. Vongkulbhisal, and A. S. Mar, “Data- efficient framework for real-world multiple sound source 2D localization,” in IBM, 2021, pp. 1–6.

[50] T. Rodemann, M. Heckmann, F. Joublin, C. Goerick, and B. Schölling, “Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping,” IEEE Int. Conf. Intell. Robot. Syst., no. Iros, pp. 860–865, 2006, doi: 10.1109/IROS.2006.281738.

[51] Y. Hu, X. Sun, L. He, and H. Huang, “A generalized network based on multi-scale densely connection and residual attention for sound source localization and detection,” J. Acoust. Soc. Am., vol. 151, no. 3, pp. 1754–1768, 2022, doi: 10.1121/10.0009671.

[52] K. Guirguis, C. Schorn, A. Guntoro, S. Abdulatif, and B. Yang, “SELD-TCN : Sound Event Localization & Detection via Temporal Convolutional Networks,” in Embedded AI: Acoustic Perception, 2020. doi: http://dx.doi.org/10.23919/Eusipco47968.2020.9287716.