BeBeC-2020-D22


# A NEURAL NETWORK BASED APPROACH TO GRIDLESS SOUND SOURCE IDENTIFICATION

Paolo Castellini, Nicola Giulietti, Nicola Falcionelli, Aldo Franco Dragoni and Paolo Chiariotti
Università Politecnica delle Marche
Via Brecce Bianche, 60131, Ancona, Italy

**Abstract**

Deep learning and Neural Networks strategies have become very popular in the last year as tools for image and data processing. In acoustics, neural network-based approaches have been typically used to recognize audio patterns, but more recently some authors applied deep learning to localize multiple-sources exploiting the grid-based approach typical of sound source localization methods or to filter/improve acoustic maps obtained by more traditional techniques like conventional beamforming. This paper wants to propose the use of artificial neural networks (ANNs) for identifying (localization and quantification) multiple sound sources in a grid-less way. The approach uses the microphones Cross-Spectral-Matrix (CSM) as input to the network and provides as output both the location and strength of sources contributing to the acoustic field. The grid-less strategy targets improving spatial resolution and computational efficiency. The proposed solution is discussed here just on simulated data for assessing its accuracy and sensitivity.

## 1 INTRODUCTION

Acoustic imaging has represented an important branch of acoustics since the '70s, in which the first beamforming algorithms was applied to this field. Since then, different algorithms and approaches have been developed. Their level level of complexity has also increased, benefiting from the improvement in data acquisition and computer computation performances. A quite comprehensive review of these techniques is presented in [5] where the main beamforming algorithms dealing with the source identification, starting from the very basics and progressing to more advanced concepts and techniques, are presented, also reporting practical examples referring to different applications. In [13] a review of the most well-known and state-of-the-art acoustic imaging methods are presented; however, the focus there is mainly on aeroacoustic applications.

No matter the approach addressed, all acoustic imaging methods are based on direct/inverse relations between microphones of the array and target points of potential sources located on a grid. The spacing between these points also identifies the accuracy in identifying the locations of the noise sources. This means that the true location of sources is highly dependent on the grid design. The method proposed in this paper is a first attempt to overcome this limit. With the intention of exploiting hidden patterns and regularities in Cross Spectral Matrices, this work proposes a neural-network-ensemble methodology for estimating both positions and strength of sound sources in a grid-less approach. This method is particularly targeted to those applications in which fixed microphone array installations are used, e.g. aeroacoustic testing. Indeed, once training of the neural-network model is performed, the identification of source locations and strengths can be performed in quasi-real time with accuracy comparable to the one of deconvolution or inverse approaches, which contrarily usually need long computing time.

As for the use of neural networks and deep learning in acoustics, despite several papers have been issued (see, for instance, [3, 6, 17, 20, 21], just to cite some), very few relates to acoustic imaging. Indeed, just few examples can be found in literature. In [9] Kujawski examines whether the use of deep neural networks can lead to an accurate characterization of single point sources from microphone array data. Starting from conventional beamforming maps, the proposed method filters out the map in order to extract the source location with sub-grid accuracy. The source coordinates are thus obtained, together with their respective strengths. The application takes advantage from the residual network architecture, a well-established model in the field of image recognition.

In [4] Chen proposes a two-step method for real-time multiple-source direct localization by modular neural network. In this method, the area of interest is divided into multiple sub-areas and Multi-Layer Perceptron (MLP) neural networks are employed to detect the presence of a source in a sub-area and filter sources in other sub-areas, while radial basis function (RBF) neural networks carry out the position estimation.

In [11] the first example in which a neural network approach is used to directly process microphone array data for acoustic mapping is presented. The complex Cross Spectral Matrix is fed to a convolutional neural network (CNN) and the training is performed considering the source distribution as the output. There is no need of providing any propagation function and microphone positions in advance, nor any knowledge of the physical meaning of the experiment. Although sidelobes may appear in some situations, the proposed technique takes advantage from the very high computing speed with respect to traditional methods. Even if the idea might be promising, their results are yet unpractical, presenting output maps with a very rough resolution of a $10 \times 10$ point grid.

## 2 MATERIAL AND METHODS

The neural network model proposed in this paper is based on a Multi-Layer Perceptron approach targeted to regression: given a set of input-output continuous variables, the task of the model is to predict new continuous outputs given new statistically independent input data.

The basic component of neural networks is the artificial neuron. This is a simple operation unit that has weighted input signals $A_1, A_2, ..., A_N$ and bias $\theta$, and produces an output signal $u$ through the activation function $f(v)$. All the weights and bias are summed and given as input for the activation function, according to equation (1):

$$u = f(\sum_{j=0}^{N} W_j * A_j + \theta) \tag{1}$$

where $W_j$ is the $j$-th input weight. The activation function can be linear or non-linear and determines the output of the artificial neuron. Commonly, non-linear activation functions are used in order to combine the inputs in more complex ways. The activation function fixes the output value boundaries and determines the neuron activation threshold. The choice of the activation function is based on the type of problem that is being model led. Then all the neurons are arranged into layers of neurons and multiple layers are arranged into a neural network. The weights, that are usually randomly initialized, are trained by a back–propagation algorithm, which is a supervised learning technique [15, 16, 18].

The model proposed in this paper, shown in fig. 1, is based on a MLP architecture with six hidden layers, with Rectifier Linear Unit (ReLU) and Linear activation functions.

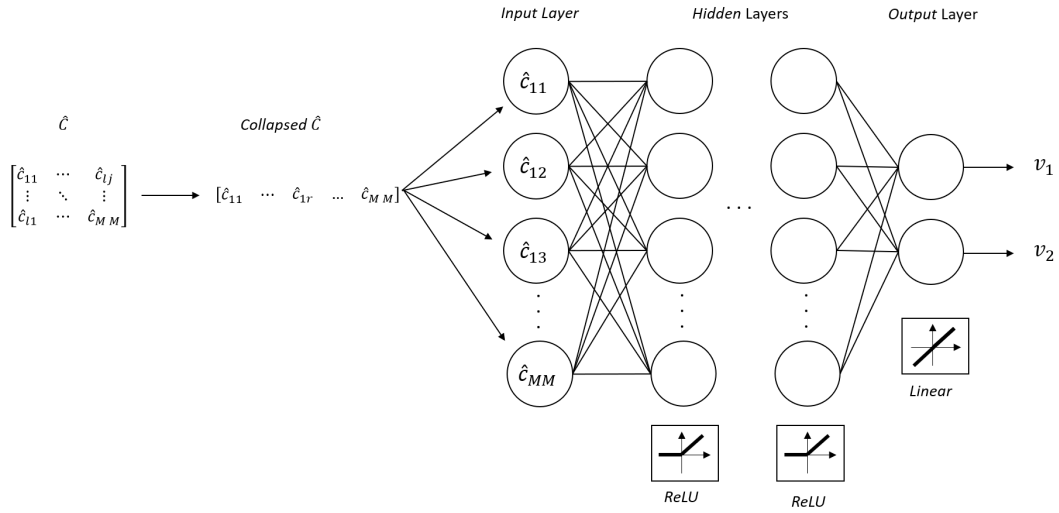


*Figure 1: Multi-Layer Perceptron model*

The input to the network is a modified Cross Spectral Matrix (CSM). The CSM matrix $\mathbf{C}$ is at first rearranged to avoid redundancy and then collapsed into a one dimensional array. Being Hermitian in nature, and being the main diagonal usually removed in aeroacoustic applications (towards which this algorithm is particularly targeted) as it contains microphones self-noise, the CSM can can be transformed in a new square matrix $\hat{\mathbf{C}} \in \mathbb{R}$ of $M \times M$. This is obtained as represented graphically in fig. 2. In fact, the CSM is split in its real and imaginary parts and then the two parts combined to create a new matrix $\hat{\mathbf{C}}$ organized as follows: the upper triangular part of $\Re(\mathbf{C})$ becomes the upper triangular part of $\hat{\mathbf{C}}$, while the upper triangular part of $\Im(\mathbf{C})$ becomes the lower triangular part of $\hat{\mathbf{C}}$. The main diagonal is set to zero.

Location and strength (amplitude and phase) of the acoustic sources are given as outputs for training the model. The ReLU activation function is given by equation (2) and even with a domain ranging from $-\infty$ to $+\infty$, the output can not assume negative values. ReLU usually helps the model learning non-linear interactions and effects, and several works demonstrate

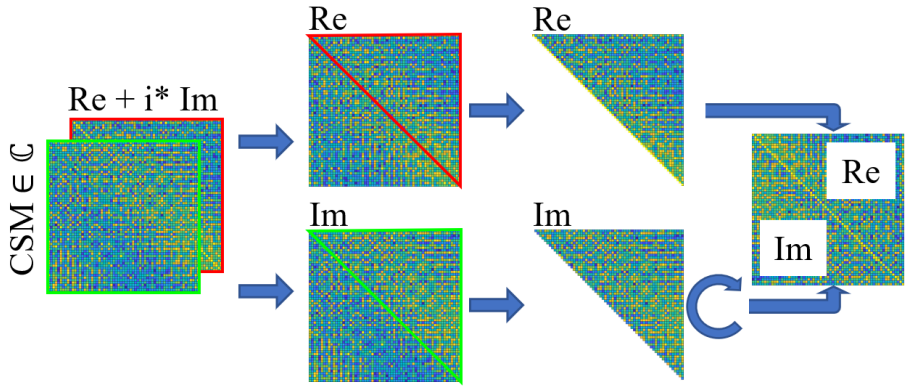


*Figure 2: Pre–processing scheme.*

significant gains in final system accuracy and training efficiency [7, 12, 19].

$$f(v) = \begin{cases} 0, & \text{for } v < 0 \\ v, & \text{for } v \geq 0 \end{cases} \tag{2}$$

Since the problem is modeled for regression, outputs are unbounded, so a Linear activation function for the output layer is chosen. The Linear activation function is given by equation (3) and ranges from $-\infty$ to $+\infty$, consequently the output can assume any value.

$$f(v) = v \tag{3}$$

The input CSM $\hat{\mathbf{C}}$ is standardized according to equation (4), where $c_{kl}$ is the $k,l$ element of $\hat{\mathbf{C}}$, and $\mu$ and $\sigma$ are respectively $\hat{\mathbf{C}}$ mean and standard deviation, in order to have a mean of 0 and a standard deviation of 1. Standardizing inputs is useful when non–linear activation functions are applied and it helps avoid getting stuck in local optimal points [1, 10].

$$\hat{c}'_{kl} = \frac{\hat{c}_{kl} - \mu_{\hat{\mathbf{C}}}}{\sigma_{\hat{\mathbf{C}}}} \tag{4}$$

Five different models are trained in order to predict:

- $(x_1, y_1)$, the location of the first strongest sound source.
- $(x_2, y_2)$, the location of the second strongest sound source.
- $q_2$, the module of the strength of the second strongest sound source.
- $(x_3, y_3)$, the location of the third strongest sound source.
- $q_3$, the module of strength of the second strongest sound source.

Mean Squared Error is used as loss function, which is defined as:

$$L(u-\hat{u}) = \frac{1}{\hat{N}} \sum_{j=0}^{\hat{N}} (\hat{u_j} - u_j)^2 \tag{5}$$

where $\hat{u_j}$ and $u_j$ are the predicted and simulated values of the $j$-th output and $\hat{N}$ is the total number of simulation performed.

The performance of the approach is tested on a simulated data set. The data set was created considering $M = 64$ microphones arranged to form a Voegl spiral according to the following equation (polar co-ordinates):

$$\begin{aligned} r &= R\sqrt{\frac{m}{M}} \\ \phi &= 2\pi m \frac{(1+\sqrt{V})}{2} \end{aligned} \tag{6}$$

with $R = 0.5$ and $V = 5$. The variable $m$ represent the $m$-th microphone of the array.

One million of cases were simulated with three sound sources, all emitting at 4 kHz, in each case. The location of these sources was varied in the range [-0.5 m; +0.5 m] in $x$ and $y$ coordinates to comply with a uniform random distribution. The strength of the three sound sources was also normalized with respect to the source with the maximum strength and varied to have uniform distribution in a dynamic range of 20 dB. Acquisition noise at microphone locations was simulated by considering additive and multiplicative noise [2, 14] as a certain Signal-to-Noise Ratio (SNR). The simulated input data were then split into Training, Test and Validation sets in a ratio of 8:1:1 for the training phase. Weights were initialized randomly. The batch size was set to 5000 and the number of epochs to 50. The Adam optimizer was used with default settings as reported in [8].

## 3 RESULTS

The five MLP models obtained from the training process of the simulated data were validated with a further set of simulated data ($N = 100000$), statistically independent from the data-sets used for the training phase.

Fig. 3 shows the loss curves of training and validation for the five models considered in terms of Mean Squared Error over epochs. It can be seen that good convergence is obtained for the models within the epoch's range adopted.

The statistical distributions of the location errors for the three sources, as well as the errors in terms of sound power levels $L_W$ (obtained from source strengths) for the second and third source, are reported in Fig. 4. The errors in sound power levels are reported in terms of dB, since they are calculated as ratio between the sound power of the current source and the sound power of the strongest source (acting as dB reference). It is well evident the Gaussian nature of the distributions, the centering around zero mean as well as the absence of skewed behaviors.

To further prove the efficacy of the approach proposed, Table 1 also reports the average values and the standard deviations of the distributions of Fig. 4. The standard deviations related to the error locations increase for the second and and third source up to approximately $1.5\lambda$ and $2\lambda$.

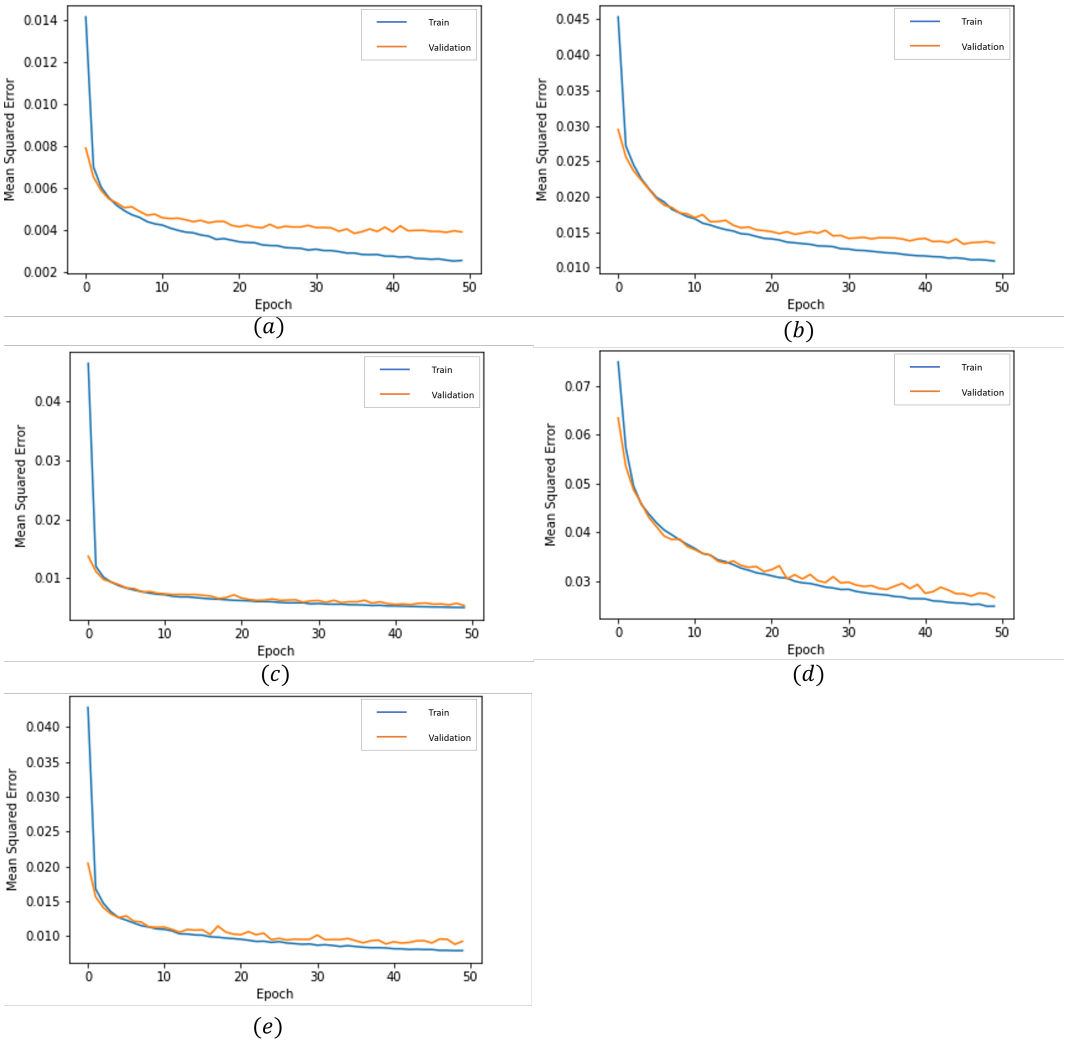


*Figure 3: Simulated test case: Model Loss curves over epochs - first source position (a); second source position (b); second source strength (c); third source strength (d); third source position (e).*

*Table 1: Simulated test case: prediction Errors average and standard deviation values for the three sources*

| | Source 1 | | Source 2 | | | Source 3 | | |
|---|---|---|---|---|---|---|---|---|
| | $x[m]$ | $y[m]$ | $x[m]$ | $y[m]$ | $L_W[dB]$ | $x[m]$ | $y[m]$ | $L_W[dB]$ |
| Average | 0.005 | -0.001 | 0.006 | -0.004 | -0.03 | -0.005 | -0.015 | -0.09 |
| Std. Dev. | 0.079 | 0.083 | 0.129 | 0.135 | 0.73 | 0.189 | 0.189 | 2.29 |

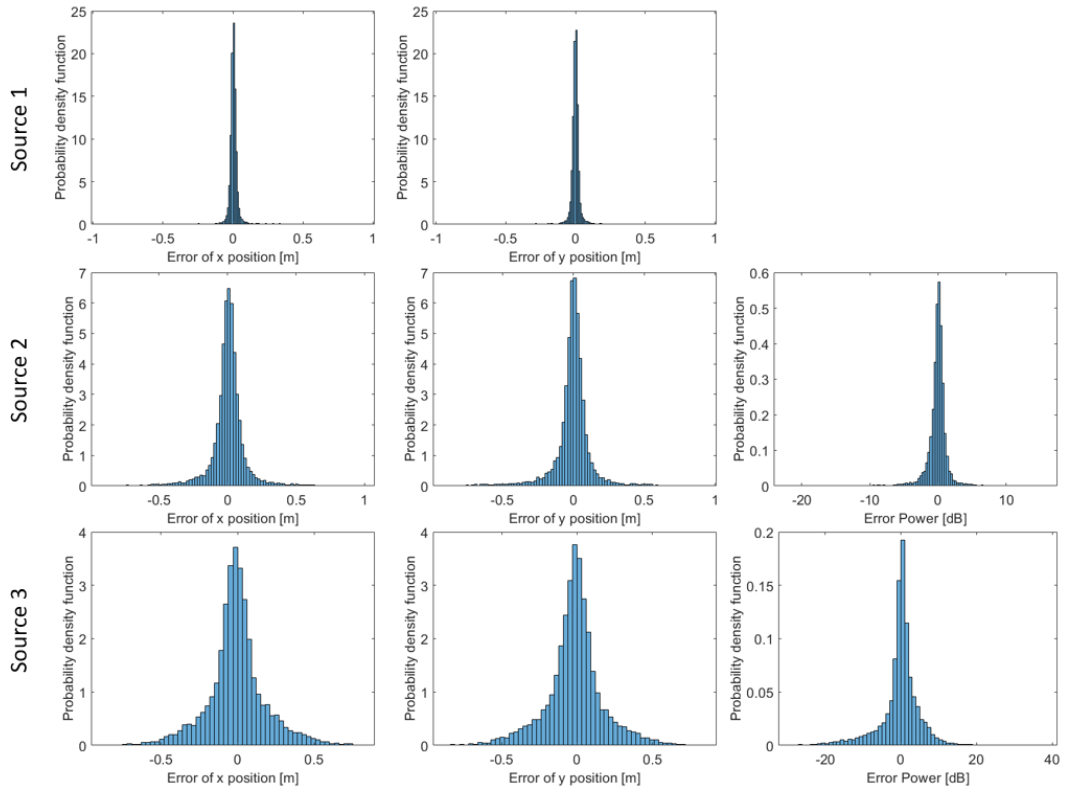


*Figure 4: Simulated test case: Histograms of prediction errors for the three sources*

## 4 CONCLUSIONS

This paper presented a novel neural-network based grid-less Sound Source Localization approach. The neural model receives as input the Cross-Spectral Matrix associated to the $M$ microphones of the array, once it is re–arranged to a non-redundant, real matrix ($M \times M$ in size).

The neural approach is based on MLP class, and provides as output the locations of multiple sources and the strengths of the sources with respect to the strongest one. The performance of the whole approach was discussed on simulated data. Sources were located with great accuracy as well as the strengths of the weaker sources were well identified with respect to the one of the strongest source.

The aim of the paper was to present a preliminary study on this novel approach, and further tests are surely needed to prove its applicability in real world scenarios. However, we think this method could be helpful in those experimental conditions in which the same array arrangement is used, like in aeroacoustic wind tunnel testing. Once the models are identified, data processing, given the grid-less nature of the method, is extremely fast and can pave the way to real-time acoustic imaging.

## REFERENCES


[1] H. Anysz, A. Zbiciak, and N. Ibadov. "The influence of input data standardization method on prediction accuracy of artificial neural networks." *Procedia Engineering*, 153, 66–70, 2016. doi:10.1016/j.proeng.2016.08.081.

[2] G. Battista, P. Chiariotti, and P. Castellini. "Spherical harmonics decomposition in inverse acoustic methods involving spherical arrays." *Journal of Sound and Vibration*, 433, 425 – 460, 2018. ISSN 0022-460X. doi:https://doi.org/10.1016/j.jsv.2018.05.001. URL http://www.sciencedirect.com/science/article/pii/S0022460X1830275X.

[3] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle. "Machine learning in acoustics: theory and applications." *The Journal of the Acoustical Society of America*, 146(5), 3590–3628, 2019. doi:10.1121/1.5133944.

[4] X. Chen, D. Wang, J. Yin, and Y. Wu. "A direct position-determination approach for multiple sources based on neural network computation." *Sensors (Basel, Switzerland)*, 18, 2018. doi:10.3390/s18061925.

[5] P. Chiariotti, M. Martarelli, and P. Castellini. "Acoustic beamforming for noise source localization – reviews, methodology and applications." *Mechanical Systems and Signal Processing*, 120, 422–448, 2019. doi:10.1016/j.ymssp.2018.09.019.

[6] A. Czyzewski. "Automatic identification of sound source position employing neural networks and rough sets." *Pattern Recognition Letters*, 24(6), 921 – 933, 2003. ISSN 0167-8655. doi:https://doi.org/10.1016/S0167-8655(02)00204-0.

[7] K. He, X. Zhang, S. Ren, and J. Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015. doi:10.1109/iccv.2015.123. URL http://dx.doi.org/10.1109/ICCV.2015.123.

[8] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization." In *3rd International Conference for Learning Representations, San Diego, 2014*. 2014. URL http://arxiv.org/abs/1412.6980.

[9] A. Kujawski, G. Herold, and E. Sarradj. "A deep learning method for grid-free localization and quantification of sound sources." *Journal of the Acoustical Society of America*, 146(3), EL225–EL231, 2019. doi:10.1121/1.5126020.

[10] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. "Efficient backprop.", 1998.

[11] W. Ma and X. Liu. "Phased microphone array for sound source localization with deep learning." *Aerospace Systems*, 2(2), 71–81, 2019. ISSN 2523-3955. doi:10.1007/s42401-019-00026-w.

[12] A. L. Maas, A. Y. Hannun, and A. Y. Ng. "Rectifier nonlinearities improve neural network acoustic models." In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. 2013.

[13] R. Merino-Martínez, P. Sijtsma, M. Snellen, T. Ahlefeldt, J. Antoni, C. J. Bahr, D. Blacodon, D. Ernst, A. Finez, S. Funke, T. F. Geyer, S. Haxter, G. Herold, X. Huang, W. M. Humphreys, Q. Leclère, A. Malgoezar, U. Michel, T. Padois, A. Pereira, C. Picard, E. Sarradj, H. Siller, D. G. Simons, and C. Spehr. "A review of acoustic imaging methods using phased microphone arrays." *CEAS Aeronautical Journal*, 10(1), 197–230, 2019. ISSN 1869-5590. doi:10.1007/s13272-019-00383-4.

[14] A. Pereira. *Acoustic imaging in enclosed spaces*. Ph.D. thesis, INSA de Lyon, 2014.

[15] F. Rosenblatt. "Principles of neurodynamics: Perceptrons and the theory of brain mechanisms." *Spartan Books*, 1961.

[16] D. E. Rumelhart and J. L. McClelland. "Learning internal representations by error propagation." In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (edited by M. P. Cambridge), chapter 8, pages 319–362. MIT Press Cambridge, 1986.

[17] D. Salvati, C. Drioli, and G. L. Foresti. "On the use of machine learning in microphone array beamforming for far-field sound source localization." In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. 2016. ISSN null. doi:10.1109/MLSP.2016.7738899.

[18] H. Trevor, T. Robert, and F. Jerome. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[19] A. Venkitaraman, A. M. Javid, and S. Chatterjee. "R3net: Random weights, rectifier linear units and robustness for artificial neural network." *aRxIV*, 2018.

[20] J. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa. "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates." *Sensors*, 18(10), 3418, 2018. doi:doi:10.3390/s18103418.

[21] X. Xia, R. Togneri, F. Sohel, Y. Zhao, and D. Huang. "A survey: neural network-based deep learning for acoustic event detection." *Circuits, Systems and Signal Processing*, 38(8), 3433–3453, 2019. ISSN 0278-081X. doi:10.1007/s00034-019-01094-1.