A Neural Network based microphone array approach to grid-less Noise Source Localization

Paolo Castellini, Nicola Giulietti, Nicola Falcionelli, Aldo Franco Dragoni, Paolo Chiariotti*

Università Politecnica delle Marche, Via Brecce Bianche 12,60131, Ancona, Italy

Abstract

Deep learning and Neural Networks strategies have become very popular in the last year as tools for image and data processing. As for acoustics, neural network-based approaches have been typically used to recognize audio patterns or features or to spatially localize a single emitting source like a speaker. More recently, some authors used 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 localizing and quantifying 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 on simulated data for assessing its accuracy and sensitivity. Preliminary investigations on real data are also reported.

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

*Corresponding author
Email address: p.chiariotti@univpm.it (Paolo Chiariotti)

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 [1] where the main beamforming algorithms dealing with 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 [2] a review of the most well-known and state-of-the-art acoustic imaging methods is also 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 best accuracy in identifying the locations of the noise sources. This means that the true location of sources is highly dependent on the grid design. There have been examples of approaches trying to overcome this limit. Malgoezar et al. [3] suggest the use of Differential Evolution (DE) with two alternative energy functions (a Bartlett processor and a cross-spectral matrix based formulation) to tackle the problem. The method proved to be quite efficient in the cases considered by the authors: the Bartlett energy function seemed to have better performance even though it has the disadvantage of not determining the source strength. Liua and Bolton [4] proposed an ESM-based (Equivalent Source Model) approach in which a non-linear optimization was performed to identify source locations and strengths that achieved the least discrepancy between the measured and the model-predicted sound fields. Indeed, the ES model characterizing the tensor approach discussed in their paper allows the component locations of the sources to be undetermined and allows different sources to have different locations. Liua and Bolton demonstrated the efficiency of their approach in reconstructing the sound field of a loudspeaker tested in anechoic conditions. The present paper wants to follow the same philosophy of the previous papers, i.e. avoid the necessity of working with grids, by exploiting the non-linear modeling capabilities of neural approaches. Even though the use of neural-based approaches in acoustics is not new, with literature reporting examples addressing the identification of direction of incoming acoustic waves in presence of noise and reverberation [5], the estimation of the three-dimensional location of an acoustic source from the raw audio signal [6] (see [7] for a survey of recent advances of machine learning and deep learning in acoustics), there are fewer documents reporting the use of microphone arrays and neural approaches for localizing multiple acoustic sources. In [8], for instance,

a weighted minimum variance distortion-less response (WMVDR) algorithm is proposed for far-field sound source localization in noisy environment. The broadband beam-forming is computed in the frequency-domain and a machine learning method is used for selecting only the narrow-band components that positively contribute to the broadband fusion. 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 [10] 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 10×10 point grid. 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. A novel arrangement of the the microphones Cross-Spectral Matrix is proposed as input to the neural model. This new arrangement aims at preserving the real and imaginary part of the CSM in a more compact formulation. The method presented in the paper is particularly targeted to those applications in which fixed microphone array installations are used (e.g. industrial installations), as the training phase can be performed just once, while the extraction of source locations and strengths can take place in quasi-real time

with accuracy comparable to the one of deconvolution or inverse approaches, whose computing time is highly related to the settings of the investigation grid. The paper is organized as follows: section 2 discusses the approach proposed, while section 3 reports results from both synthetic and experimental datasets; section 4 draws the main conclusions of the work.

2. MATERIAL AND METHODS

2.1. Data-set creation

When dealing with Neural-based techniques it is of quite importance to collect significant amount of data to be used in the training, validation and, eventually, testing phase of the model. To achieve this goal, a "synthetic data generator" has been developed. The code was designed to perform the following tasks:

- generate/load microphone arrays of various design (e.g. random, spiral, etc.) with M microphones;
- generate a set of monopole sources (either fully/partially correlated or uncorrelated), each s-th source characterized by:
  - random location $r_s(x, y)$ on a given area at a given distance from the array;
  - random strength Q_s.
- propagate the sources from their locations to each q-th microphone of the array, so as to generate a complex vector of noise-free pressure $\hat{\mathbf{p}}$ at microphone locations;
- simulate acquisition noise at microphone locations; this is done by considering additive and multiplicative noise at a certain Signal-to-Noise Ratio (SNR). The noise term on each microphone is calculated, according to [12, 13], as

$$z_q = 10^{-SNR/20} \left(\gamma \mathrm{e}^{\mathrm{i}\varepsilon} \hat{p}_q + \delta \mathrm{e}^{\mathrm{i}\zeta} \sqrt{\frac{\|\hat{\mathbf{p}}\|^2}{M}} \right), \tag{1}$$

  where γ and δ are zero mean Gaussian random variables ($Var(\gamma) = Var(\delta) = 1$), ε and ζ are random variables uniformly distributed between 0 and 2π. The complex noise vector is $\mathbf{z}$ and the total pressure

is $\mathbf{p} = \hat{\mathbf{p}} + \mathbf{z}$.

- order the sources in descending strength order;
- express the phase of each source as phase difference with respect to the strongest source;
- generate the Cross Spectral Matrix ($\mathbf{C}$) for the microphone array adopted, in accordance to (2);

$$\mathbf{C} = \mathbb{E}\left(\mathbf{p}\mathbf{p}^{\mathrm{H}}\right) \tag{2}$$

where $\mathbb{E}$ is the expectation operator working on Ξ realizations of the complex pressure vector $\mathbf{p}$, whose elements are the noisy complex pressures $P_m(\omega_k)$ provided at each microphone location, and the superscript $^{\mathrm{H}}$ represents the complex conjugate transpose operator.

Given the M microphones, S sources at a given frequency and a population of N random sets of data, a 3D-complex CSM matrix of $M \times M \times N$ in size can be obtained. The corresponding actual set of noise sources is stored in a 3D matrix of $4 \times S \times N$ elements, being each source defined with 2 position coordinates, the strength and the phase.

2.2. Data pre-processing

The main idea underling the approach proposed in this paper grounds on the consideration that acoustic sources at given locations, and of given strengths, combine to provide an image of the sound field that is reflected in the microphones CSM. Under this assumption, the CSM itself can be considered as an image. Hence, a neural model can be trained to identify modifications to this image, thus relating the CSM to source locations and strengths. Keeping this "image-like" representation of the CSM, it is clear that the CSM should be somehow rearranged to be digested by neural models. Indeed, when dealing with images, neural models can either work with color images (hence three-dimensional matrix whose third dimension carries the color planes - e.g. in a RGB representation) or grayscale images (two-dimensional matrix representation). To be represented in an "image-like" arrangement, the CSM $\mathbf{C}$ should be rearranged to suit one of the former representation avoiding redundancy. The CSM is Hermitian in nature; moreover, in many applications the main diagonal carries microphones self-noise (this is because each individual channel can be contaminated by inchoerent noise,

and the contamination will be concentrated on the diagonal of the CSM if assuming stationary signals and long-time averaging) hence should be removed. The CSM $\mathbf{C}$ can then be transformed in a new two-dimensional square matrix, $\hat{\mathbf{C}} \in \mathbb{R}$ of $M \times M$, by combining its real and imaginary parts 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. This is represented graphically in fig. 1. The input CSM $\hat{\mathbf{C}}$ should then be standardized according to equation (3), where c_{kl} is the k, l element of $\hat{\mathbf{C}}$, and μ and σ 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 [14, 15].

$$\hat{c'}_{kl} = \frac{\hat{c}_{kl} - \mu_{\hat{\mathbf{C}}}}{\sigma_{\hat{\mathbf{C}}}} \tag{3}$$

Standardizing the CSM does not alter the geometric information about source locations carried by the CSM. However, it is clear that it prevents the possibility to predict the absolute strength of each source. However, the relative strength ratios with respect to the strongest source are kept, given the linear nature of the standardization operation.

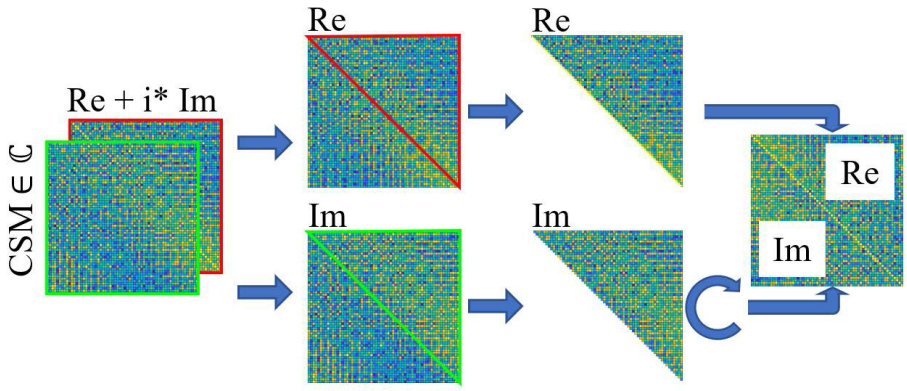

Figure 1: CSM Pre–processing: rearrangement of CSM to move from complex-valued to real-valued data to be used as input to the neural model.

2.3. Multi-Layer Perceptron model

The neural network model proposed in this paper is based on a Multi-Layer Perceptron (MLP) 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.

MLP is a specific class of feed-forward Artificial Neural Networks (ANNs) that has (i) an input layer, (ii) one or more hidden layers, and (iii) an output layer, which can be shaped based on the task (i.e. classification or regression in most cases). The model proposed in this paper, shown in fig. 2, is based on a MLP architecture with six hidden layers, with Rectifier Linear Unit (ReLU) and Linear activation functions. More precisely, the network layers' dimensions and types are shown in table 1, for a total number 845612 parameters[1].This number of parameters implies that N (random set of data) should be set to ensure $4 \times S \times N \gg$ network parameters. For instance, in the tests reported in this paper, a dataset of one million testing cases (N) and three sources (S) was used to comply with this condition.

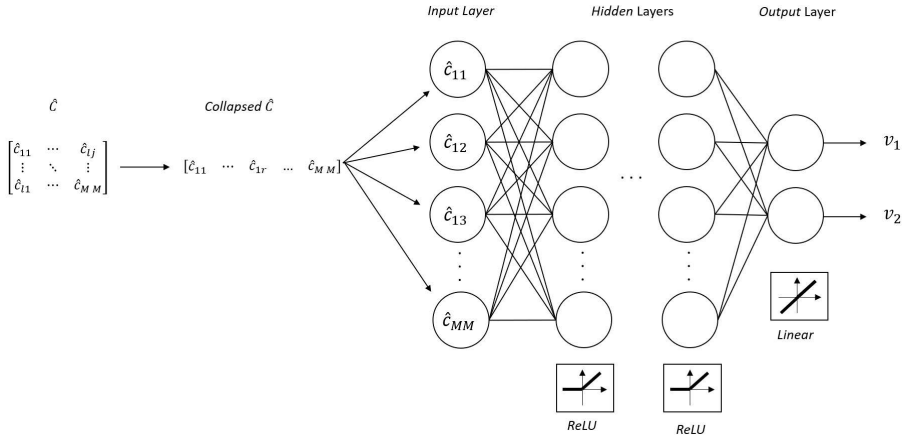

Figure 2: Collapsed CSM is mapped to Multi-Layer Perceptron's input layer

The CSM, collapsed into one dimensional array, is given as input, while location and strength (amplitude and phase – relative values with respect to the strongest source) of the acoustic sources are given as outputs for training the model. The ReLU activation function is given by equation (4) 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 significant gains in final system

[1]One weight for each connection between two neurons plus one bias for each neuron.

Layer No.	No. of Neurons	Activation Function	Type
1	4096	ReLU	Input
2	200	ReLU	Fully Connected
3	100	ReLU	Fully Connected
4	50	ReLU	Fully Connected
5	20	ReLU	Fully Connected
6	2	ReLU	Fully Connected

Table 1: Proposed approach's MLP architecture of the proposed approach.

accuracy and training efficiency [16–18].

$$f(u) = \begin{cases} 0, & \text{for } u < 0 \\ u, & \text{for } u \geq 0 \end{cases} \tag{4}$$

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 ranges from $-\infty$ to $+\infty$, consequently the output can assume any value.

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 relative to the strongest one.
- (x_3, y_3), the location of the third strongest sound source.
- q_3, the module of strength of the third strongest sound source relative to the strongest one.

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

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

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

2.4. Training, Validation and Test data sets

Two different data sets were investigated to prove the performance of the approach. One data set is a synthetic data set, while the second one is related to a real experiment performed in semi-anechoic conditions.

2.4.1. Synthetic data set

The synthetic data set was created considering $M = 64$ microphones arranged to form a Vogel spiral [19] 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 $V = 5$.

One million of CSM cases, in which each CSM $\mathbf{C_n}$ ($n = 1...N$) was obtained over $\Xi = 100$ realizations, were synthesized 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 (array to target distance = 2 m) 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. The synthetic 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 [20].

2.4.2. Experimental data-set

The experimental data set refers to the set-up described in [21] and shown in Fig. 3.

Two loudspeaker were installed 0.6 m far from a microphone array of $M = 43$ microphones in random arrangement. The two loudspeakers were fed with the same random noise (2 kHz-8 kHz) to have fully correlated sources. Fig. 3 (right) shows the microphones arrangement with respect to the source location projected over the array area. The MPL model was trained using a set of synthetic data ($\hat{N} = 800000$) in which three sources and the random microphones arrangement of Fig. 3 were considered. As for the synthetic

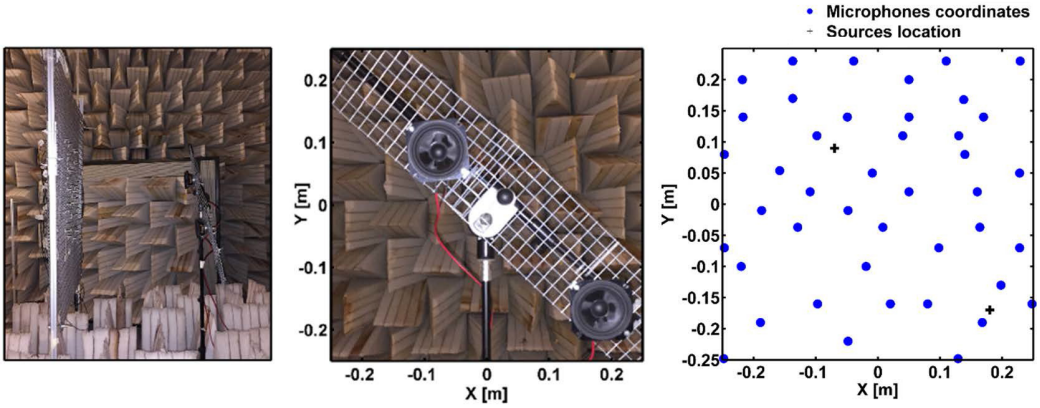

Figure 3: Experimental setup of array measurements on loudspeakers in semi-anechoic conditions: microphone array vs source locations - cones lying on the bottom were added for the test, but they are not a fixed installation (left); source locations (centre); microphones and sources distribution (right)

data set of Section 2.4.1, the location of the three sources was varied within the range [-0.5 m; +0.5 m] in x and y coordinates to generate a uniform random distribution of potential source locations over the investigation area. The same dynamic range was also adopted in creating the source strength distributions.

The aim of this test was to prove the performance of the approach in case training is performed with a number of sources differing from the actual number of sources and to demonstrate the absence of dependencies to array configurations, given the choice of a random array with respect to a spiral-arranged one. Moreover, it is well known that the behavior of real loudspeakers differs from that of ideal monopoles used in simulated data (e.g. emissivity pattern, efficiency, etc.) thus further widening differences between simulated and real data sets.

3. RESULTS

The experiments highlight good performance in terms of accuracy and generalization capabilities of the model. Indeed, the approach shows a generally good learning behavior, avoiding common pitfalls like overfitting [22], as:

- dataset splits are completely separated;

- the dataset is composed of an high number of samples (1000000);
- training and testing losses, both in the simulated and real experiments, are comparably low.

3.1. Synthetic dataset

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

Fig. 4 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.

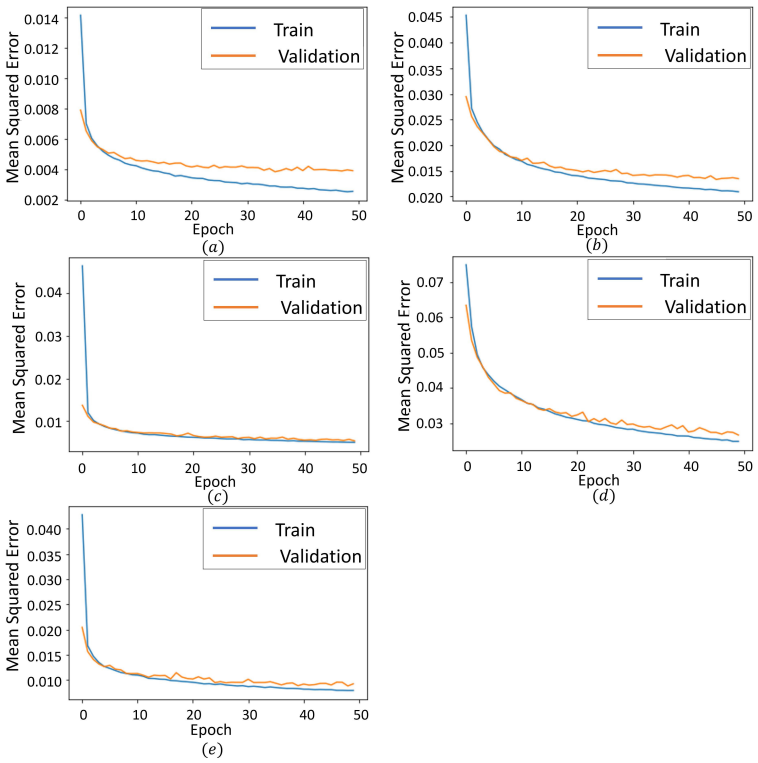

Figure 4: Synthetic dataset: Model Loss curves over epochs - first source position (a); second source position (b); second source strength (c); third source strength (d); third source position (e).

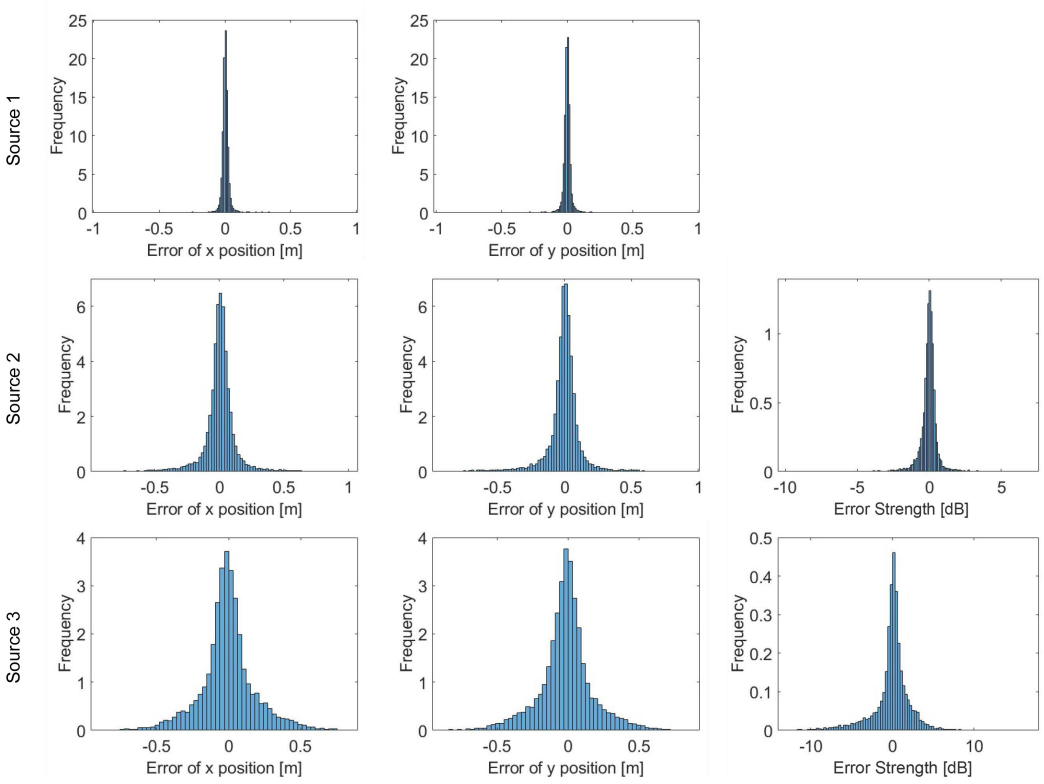

Figure 5: Synthetic dataset: Histograms of prediction errors for the three sources

The statistical distributions of the location errors for the three sources, as well as the errors in terms of strength (volume velocity) levels L_q for the second and third source, are reported in Fig. 5. The errors in strength levels are reported in terms of dB, since both predicted and ground-truth values are calculated as ratio between the strength of the current source and the strength of the strongest source (acting as dB reference). This formulation in presenting results is adopted since only strengths relative to the strongest source can be estimated by the model, given the standardization pre-processing the CSM goes through. If looking at the different distributions, it is well evident the centering around zero mean as well as the absence of skewed behaviors.

To further prove the efficacy of the approach proposed, Table 2 also reports the average values and the standard deviations of the distributions of Fig. 5. The standard deviations related to the error locations increase for the second and and third source up to approximately 1.5λ and 2λ. This

Table 2: Synthetic dataset: prediction Errors average and standard deviation values for the three sources

	Source 1		Source 2			Source 3		
	$x[m]$	$y[m]$	$x[m]$	$y[m]$	$L_q[dB]$	$x[m]$	$y[m]$	$L_q[dB]$
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

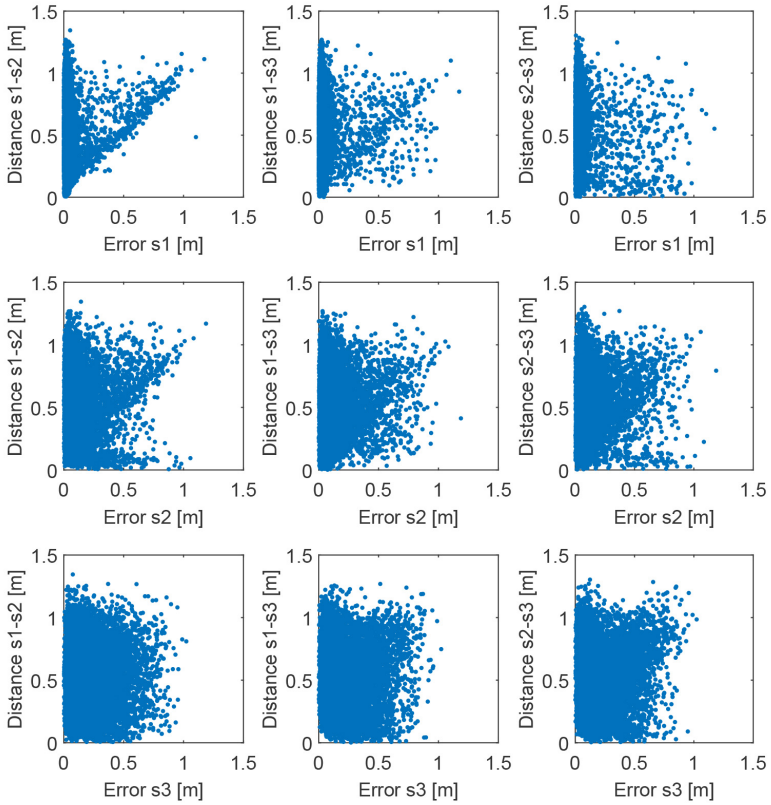

Figure 6: Synthetic dataset: Localization error sensitivity to inter-source spatial distance

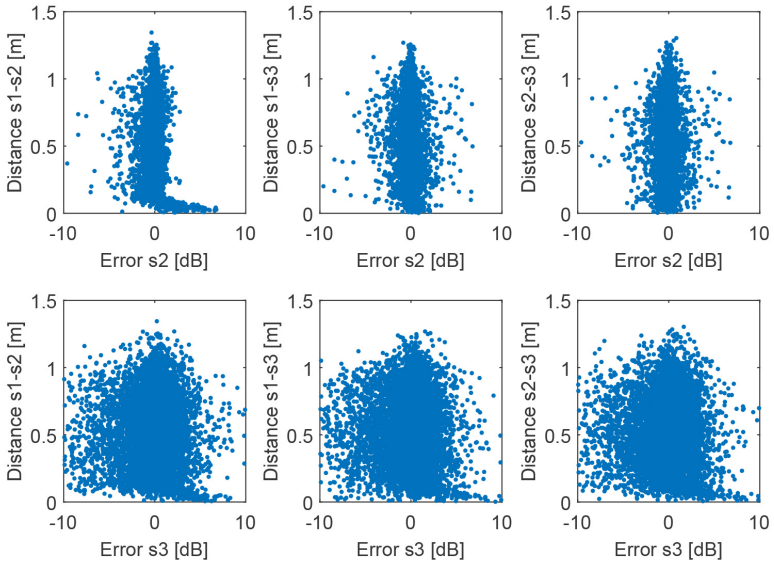

Figure 7: Synthetic dataset: Strength sensitivity to inter-source spatial distance (error expressed in dB - $\mathrm{dB}_{ref} 1m^3/s$)

suggests that weaker sources can still be localized properly (the localization is precise), but less accurately. This is also true for the relative strength of sources. Indeed, if one the one hand the average value of the error on the strength level keep the same order of magnitude moving from Source 2 to Source 3, on the other hand this does not apply to the standard deviation values. Indeed, the standard deviation value associated to the strength of Source 3 is three times higher than the one associated to the strength of Source 2. This means, more generally, that the weaker the source, the more spread can be the result in terms of source location and strength. Still, the results obtained are important, given the variability of the strength of the weak sources within a 20 dB range with respect to the strongest source. Another aspect worth to be discussed is the sensitivity of the whole approach to inter-source spatial distance. Indeed, it is interesting to understand how the approach behaves when sources are closer or when they are farther from each other. Figure 6 shows the prediction errors on the localization of each source (calculated as the distance between the predicted and the ground-truth source location) for the different inter-source distances tested in the synthetic dataset. It is quite evident that source localization errors tend to

be more important as the strength of the sources weaken, but the inter-source distance does not seem to interfere too much on the localization capability of the whole approach. As far as the strength prediction error is concerned, this is analyzed in Figure 7 with respect to inter-source distances. Contrarily to localization error, it seems that a the error on the strength prediction tends to be more important the closer are the sources to the strongest one. This is indeed quite expectable, as the strongest source may mask the presence of weaker sources the closer they are.

3.2. Sensitivity to correlation among sources

The robustness of the approach with respect to the correlation degree of sources acting in the sound field was also tested by applying the models proposed to the combination of sources *vs.* training-testing schemes addressed in Table 3. The same settings reported in Section 2.4.1 were used to gener-

Table 3: Synthetic dataset: combination of training-testing schemes for different source correlation cases

	Training	Testing
1.	Uncorrelated sources	Uncorrelated sources
2.	Fully correlated sources	Uncorrelated sources
3.	Uncorrelated sources	Fully correlated sources
4.	Fully correlated sources	Fully correlated sources

ate two different datasets involving fully correlated and uncorrelated sound sources. Hence, the main difference between the generated datasets is related to the correlation degree among the three sources synthesized. The results of the four combinations reported in Table 3 are provided in Table 4 in terms of global Mean Square Error for the five models developed. In this table, the *Position* label refers to the euclidean distance calculated between the location of the source identified by the model with respect to the ground truth of that source, while the *Strength* label addresses the strength of a source with respect to the strongest. It is interesting to notice that MSE values are almost the same for all the four combinations addressed. This is particularly interesting, as it suggests that the approach proposed is robust to any misinterpretation of correlation levels among the sources.

Table 4: Synthetic dataset: MSE associated to the combination of training-validation schemes for the different source correlation cases synthesized

	$\text{Position}_{\text{Source 1}}$	$\text{Position}_{\text{Source 2}}$	$\text{Position}_{\text{Source 3}}$	$\text{Strength}_{\text{Source 2}}$	$\text{Strength}_{\text{Source 3}}$
1.	0.004	0.014	0.027	0.004	0.007
2.	0.004	0.013	0.027	0.004	0.008
3.	0.004	0.014	0.027	0.004	0.007
4.	0.004	0.013	0.027	0.004	0.008

3.3. Experimental dataset

Results reported hereafter refer to a frequency of 2kHz; the two sources are expected to have a strength level $L_q = -50.6$ dB ($\text{dB}_{ref} 1m^3/s$) each (the speakers are fed with the same random noise, as reported in 2.4.2).

Figure 8 shows the results obtained with the proposed approach ($*$) with respect to the actual location of the speakers (dashed circles). The results obtained performing, on the same dataset, CLEAN-SC on two different calculation grids (5 mm,$\circ$, and 1 mm,+) are also reported in the same figure. Both the three approaches seem to identify quite well the two loudspeaker, even though it seems that CLEAN-SC tends to localize the sources towards the edge of the loudspeakers' cone. It should be highlighted that, for easing the comparison among the approaches, data of CLEAN-SC refer to the maximum values retrieved by the method.

It should be also recalled that the neural model infers the strength of all sources with respect to the strongest. To retrieve the strength, a steepest descent optimization algorithm was adopted. The algorithm uses the locations of the sources identified by the neural model and builds synthetic CSMs assigning different strengths to the strongest source. Since the relative strengths of the other sources with respect to the strongest are known, it is possible to synthesize a CSM by propagating all sources at microphone locations. The cost function in the optimization process is the least square error between the synthesized CSM and the measured CSM. Optimization stops when this error is minimized. It should also be highlighted that, since the relative strengths of the sources with respect to the strongest are known, the optimization step reduce to a single parameter optimization that can be run easily and fast. In the following analysis, a free-field analytical propagation model was adopted.

The accuracy in reconstructing the source locations and strengths can be further proved looking at Table 5.

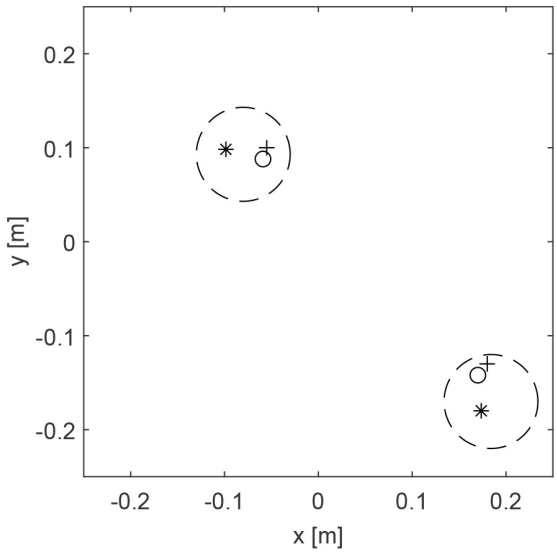

Figure 8: Experimental dataset: actual loudspeaker locations (dashed circles); source locations from: neural approach (∗), CLEAN-SC on grid with 5 mm spacing (∘), CLEAN-SC on grid with 1 mm spacing (+)

Table 5: Experimental dataset: source localization and strength (volume velocity) prediction Errors - $\mathrm{dB}_{ref} 1m^3/s$

	Source 1			Source 2			Source 3		
	$x[m]$	$y[m]$	$L_q[dB]$	$x[m]$	$y[m]$	$L_q[dB]$	$x[m]$	$y[m]$	$L_q[dB]$
Actual source	-0.080	0.090	-50.6	0.180	-0.170	-50.6	/	/	/
NN-approach	-0.098	0.098	-50.5	0.173	-0.179	-50.7	0.110	0.262	-61.2
CLEAN-SC (5 mm)	-0.055	0.100	-50.6	0.180	-0.130	-51.0	/	/	/
CLEAN-SC (1 mm)	-0.059	0.088	-50.7	0.170	-0.142	-50.8	/	/	/

It can be seen that the proposed approach well reconstructs both the locations and the strengths associated to the two sources. The difference between the actual locations of the sources (considered here as the center of the loudspeakers) and the reconstructed locations ranges from 7 mm to

18 mm. Moreover, despite the use of a different array configuration, these results perfectly fit within the ranges reported in Table 2 if considering a coverage factor of 2 on the standard deviation values estimated from the error analysis of Section 3.1.This is a well acceptable error, also considering the uncertainty associated to the actual location of the noise sources on the speakers. CLEAN-SC seems to tend to localize the sources towards the edge of the loudspeakers. It is also interesting to notice that, as far as quantification is concerned, CLEAN-SC underestimated the second source with respect to the neural approach. Moreover, the strength level of the sources identified changes when different grid spacings are used. The quantification errors with respect to the first source are comparable to those of the neural approach. Indeed, if looking at strength levels, a difference of 0.1 dB with respect to the actual strength estimated for each source can be noticed on the neural approach. It is interesting that the the underestimation of 0.1 dB identified on the second source is actually reported as an overestimation of the same level on the first source. The neural approach also assigns a very small value of strength to the third source (not present in the real experiment, but present in the synthetic data set associated to the experimental test case). However, the strength level associated to this source is more than 10 dB lower than the first source. It is clearly a non-negligible value, yet a comment is needed: as for the structure of the whole method, the model is forced to look for three sources, hence the identification of a "fake" source might be expected; nevertheless, the strength of this third "fake" source is small if compared to those of the other two sources. In this sense, this is a quite remarkable result, proving that the approach is able to provide precise output even if the number of actual sources differ from the one adopted in the training phase. Moreover, spatial accuracy is improved with no limits associated to the presence of a grid. This is another important result: indeed commonly adopted algorithms like deconvolution approaches or inverse approaches, but even simpler methods like conventional beamforming, do have accuracy and computation time somehow related to the grid resolution (once the extension of the grid is fixed). The neural approach proposed in this paper overcomes this limit. In fact, computation time (currently approximately 25 ms - inference performed on a NVIDIA GEFORCE RTX 2070 + 500 ms for the optimization step) remains unaltered once the neural models are identified, since the overall approach does not ground on the use of any grid at all. As proof of this concept, the two CLEAN-SC processing ran in 21 s and 2 s respectively for the 1 mm and 5 mm grid spacing.

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 both synthetic and experimental datasets. The two test cases discussed in the paper demonstrated that the method works fine no matter the array adopted (number of microphones, array design, etc.), with no limits on the correspondence between the number of sources involved in the training phase of the models and the number of actual sources to be identified.
Indeed, the paper is to be listed among those few methods aiming at demonstrating the potential of neural approaches in array acoustics targeted to sound source localization and quantification. As such, the authors preferred to focus on test cases that could be easily interpreted and could easily show the performance of the method proposed. Nevertheless, the approach presented could be helpful in those experimental conditions in which the same array arrangement is used. This is often the case of industrial installations. In those cases, once the models are identified, the data processing phase, given the grid-less nature of the method, is extremely fast and might enable reducing operating costs. Aeroacoustic testing might be one of the applications that could benefit the most from the adoption of this approach. Of course, authors are aware that the robustness of the method to identify spatially extended sources, or even sources with more complex directivity is yet to be proven, but they are currently working in this direction, as they strongly believe that neural-based approaches will represent, in the near future, fertile investigation topics in the array acoustics field, thus playing the role that deconvolution or inverse approaches have been playing in the last decades.

REFERENCES

[1] P. Chiariotti, M. Martarelli, P. Castellini, Acoustic beamforming for noise source localization – reviews, methodology and applications, Mechanical Systems and Signal Processing 120 (2019) 422–448.

[2] 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, C. Spehr, A review of acoustic imaging methods using phased microphone arrays, CEAS Aeronautical Journal 10 (2019) 197–230.

[3] A. M. N. Malgoezar, M. Snellen, R. Merino-Martinez, D. G. Simons, P. Sijtsma, On the use of global optimization methods for acoustic source mapping, The Journal of the Acoustical Society of America 141 (2017) 453–565.

[4] Y. Liua, J. S. Bolton, Sound field reconstruction using multipole equivalent source model with un-fixed source locations, The Journal of the Acoustical Society of America 144 (2018) 2674.

[5] A. Czyzewski, Automatic identification of sound source position employing neural networks and rough sets, Pattern Recognition Letters 24 (2003) 921 – 933.

[6] J. Vera-Diaz, D. Pizarro, J. Macias-Guarasa, Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates, Sensors 18 (2018) 3418.

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

[8] D. Salvati, C. Drioli, 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), pp. 1–6.

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

[10] X. Chen, D. Wang, J. Yin, Y. Wu, A direct position-determination approach for multiple sources based on neural network computation, Sensors (Basel, Switzerland) 18 (2018).

[11] W. Ma, X. Liu, Phased microphone array for sound source localization with deep learning, Aerospace Systems 2 (2019) 71–81.

[12] G. Battista, P. Chiariotti, P. Castellini, Spherical harmonics decomposition in inverse acoustic methods involving spherical arrays, Journal of Sound and Vibration 433 (2018) 425 – 460.

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

[14] H. Anysz, A. Zbiciak, N. Ibadov, The influence of input data standardization method on prediction accuracy of artificial neural networks, Procedia Engineering 153 (2016) 66–70.

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

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

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

[18] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015 IEEE International Conference on Computer Vision (ICCV) (2015).

[19] H. Vogel, A better way to construct the sunflower head, Mathematical Biosciences 44 (1979) 179–189.

[20] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference for Learning Representations, San Diego, 2014.

[21] C. Colangeli, K. Janssens, P. Chiariotti, P. Castellini, Clustering inverse beamforming for vehicles NVH, in: 24th International Congress on Sound and Vibration, ICSV 2017.

[22] B. Ghojogh, M. Crowley, The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial, 2019.

Credit author statement

Paolo Castellini: Conceptualization, Methodology, Software, Formal analysis, Visualization **Nicola Giulietti**: Methodology, Software, Visualization **Nicola Falcionelli**: Methodology, Writing- Original draft preparation. **Aldo Franco Dragoni**: Conceptualization. **Paolo Chiariotti**: Conceptualization, Methodology, Software, Formal analysis, Writing- Original draft preparation, Writing- Reviewing and Editing,