BeBeC-2022-S02

FREQUENCY DOMAIN BEAMFORMING USING NEURONAL NETWORKS

Armin Goudarzi
German Aerospace Center (DLR)
Bunsenstraße 10, 37073 Göttingen, Germany

Abstract

In our contribution, we examine the use of Neuronal Networks for beamforming in the frequency domain. While conventional beamformers and deconvolution algorithms are often a compromise of spatial resolution, assumptions about the Green's function, and computational effort, machine learning offers a new perspective on beamforming. While there exists some research that shows the potential of Neuronal Networks for beamforming, this study focuses on the different use-cases, and challenges that arise from the corresponding architectures. We discuss possible input and output designs for Neuronal Networks, different layer designs, and different activation-, and loss functions. The different architectures are evaluated and discussed for several source types such as monopoles and dipoles using several metrics for the quality of the resulting maps.

1 Introduction

Multiple noise-generating phenomena and mechanisms exist in acoustics. For the localization and estimation of the sound power of complex source geometries, such as planes, cars, or trains, beamforming is a reliable method [20]. Since for a sound field observed with a finite amount of sensors there exist an infinite amount of possible source configurations [12], beamforming methods rely on several assumptions. The main assumptions include generally one or all of these: spatially compact sources, monopole sources, incoherent sources, and independent sound radiation for each frequency. While all of these assumptions are typically violated in real-world scenarios, simple algorithms such as Conventional Beamforming (CB) are still widely popular due to their robustness and known limitations [20]. More sophisticated approaches exist, such as inverse methods [28], deconvolution methods such as CLEAN-SC [24], and DAMAS [5] where the true source distribution is reconstructed from the dirty beamforming maps, or unsupervised learning methods [2, 8, 9, 29] where beamforming is treated as a blind

source separation problem. However, inverse methods and advanced deconvolution methods are often computationally expensive and still include assumptions about the source.

In recent years supervised learning became widely popular and outperformed any other classification algorithms given enough learning data. For supervised learning in the frequency domain, the true source distribution and the corresponding Cross Spectral Matrix (CSM) must be known, which is mathematically the most compact form for quasi-stationary, proper, and circular data [1]. In this paper we will assume these properties, however, it was shown that this is not necessarily the case for microphone data obtained in wind tunnels. Then, higher-order statistics can improve the performance of the beamforming process [22].

For the supervised learning, we will employ an Artificial Neuronal Network (ANN), which is a universal function approximator based on the universal approximation theorem. The goal of the ANN is to generate the correct source distribution from the presented input CSM. While ANN research in non-acoustical time-domain beamforming has already advanced significantly, acoustical ANN beamforming is still relatively new. In this paper, we will solely focus on beamforming in the frequency domain, where different architectures have been proposed. On the input side, the ideas can be separated into methods that either use the CSM or CB maps. For the hidden ANN layers typically Fully Connected Layers (FCL) or Convolutional Layers (CL) are used. For the output either the source distribution is estimated on a grid, or the existence of a source, its strength, and its coordinates are estimated. The recent research is summed up in Table 1 and separated into the different input and output designs.

	grid-free	grid-based
CB map	Kujawski et al. [13]	Pinto et al. [21]
CSM	Castellini et al.[7] Lee et al. [16] Lee et al. [15]	Ma and Liu [19] Xu et al. [26]

Table 1: Overview of the research that has been conducted, separated according to the input and output designs.

All of the presented literature discusses the detection of compact monopoles at single frequencies. However, methods such as DAMAS and CLEAN-SC give very good results for these used cases. Thus, in this paper, we will discuss how to generate arbitrary training data for multipole sources such as dipoles, and distributed sources with any given coherence between them. For the ANN, we will only discuss approaches that are based on the input CSM, since the CB map input already discards information about the true source distribution and includes the source assumptions.

For the traditional grid-based approach we will discuss the several challenges and design options that arise from this problem and evaluate these based on error metrics, that are designed to quantify the accuracy of the classification and regression results. Additionally, we will explore a very recent advancement in machine learning as a work in progress: Deep Sets [27], where the beamforming problem can be reformulated as a tensor-to-set problem, which is

particularly successful in object detection and classification tasks [6, 14, 18].

2 Data generation

For deep learning a large quantity of training data is necessary. Thus, we will generate synthetic training data directly in the frequency domain, which allows the generation of millions of sources within seconds. We will explore how to generate compact and distributed sources, and monopoles and dipoles. The number of microphones is denoted by N and the sensor positions with x_n for $n = 1,\dots,N$. The source locations are denoted with y_{true}.

2.1 Source Definition

A source in this paper has an amplitude a, an arbitrary phase, a location y that can be spatially distributed, and a rotation (for multipoles).

2.2 Multipoles

The CSM corresponding to a multipole signal is given by the complex microphone pressure vector p

$$p = \begin{pmatrix} p(x_1) \\ \vdots \\ p(x_N) \end{pmatrix} \tag{1}$$

with

$$\mathbf{C} = pp^* \tag{2}$$

and * is the Hermitian transpose. Given a source signal $s(y)$, the complex pressure at microphone position x_n, is given by

$$p(x_n) = s(y)h(x_n, y) \ , \tag{3}$$

where h describes the propagation from source location y to microphone of the given source type. For a monopole the propagation function h is given by the Greens Function

$$h_{\text{mono}}(x_n, y) = \frac{\exp(-jkd)}{d} \ , \tag{4}$$

with the wavenumber $k = w/c$, c is the speed of sound, and $d = |x_n - y|$. For a dipole, the propagation function is given by

$$h_{\text{dip}}(x_n, y) = \left(\mathbf{e}_{\text{dip}} \cdot \mathbf{e}_n\right) \exp(-jkd) \left(\frac{1}{d^2} + \frac{jk}{d}\right) \tag{5}$$

with $\mathbf{e}$ being the normalized direction vector of the dipole and the microphone position n. If given in spherical coordinates, which makes it easier to control the rotation and strength of the dipole independently, they can be derived with

$$\mathbf{e} = [\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta)]^T \ . \tag{6}$$

Thus, the dot product of unit vectors incorporates the dipole directivity, based on the dipole rotation and locations of microphones in three-dimensional space.

2.3 Distributed sources and source coherence

In real world scenarios sources are often spatially distributed. To build a distributed source of any shape, we approximate it with a superposition of M compact sub-sources at source locations y_i

$$s = \begin{pmatrix} s_1(y_1) \\ \vdots \\ s_M(y_M) \end{pmatrix} . \tag{7}$$

The number of sub-sources should be chosen so that $\Delta|x_i| \ll \lambda$, to prevent aliasing in the resulting sound field. Using the corresponding propagation function h, the propagation matrix $\mathbf{H}$ is given by

$$\mathbf{H}_{nm} = h(x_n, y_m) \quad n = 1,\dots N \quad m = 1,\dots M . \tag{8}$$

Given that $s \in \mathbb{C}^M$ is a vector-valued random variable, its correlation matrix is $\mathbb{E}(ss^*)$ and the CSM $\mathbf{C}$ at the microphone array is given by

$$\mathbf{C} = \mathbf{H}\mathbb{E}(ss^*)\mathbf{H}^* . \tag{9}$$

Thus, the CSM values depends on the coherence of the sub-sources, for incoherent sub-sources one gets

$$\mathbb{E}(ss^*) = \operatorname{diag}\left(|s_1|^2, \dots, |s_M|^2\right) . \tag{10}$$

We want to replace the correlation matrix with an explicit expression depending on an amplitude vector $\hat{s} \in \mathbb{C}^M$ (defined later) and a coherence matrix $\Gamma \in \mathbb{C}^{M\times M}$. More precisely, we replace eq. 9 with

$$\mathbf{C} = \mathbf{H}(\Gamma \otimes (\hat{s}\hat{s}^*))\mathbf{H}^* , \tag{11}$$

where $\otimes$ denotes the pointwise (Hadamard) product. A function for Γ can be freely defined, e.g. $\Gamma = \mathbf{1}$ (matrix of ones) for coherent sub-sources or $\Gamma = \mathbf{I}$ (identity matrix) for incoherent sources. Additionally, Γ may include phase relations between the sub-sources s_i and s_j. However, Γ must be Hermitian so that the CSM is also Hermitian. For the purpose of this paper we define a simple variable coherence length L_c. Based on the sub-sources' wave-length λ and distance matrix between all sub-sources $\mathbf{d}$.

$$\Gamma = \begin{cases} \mathbf{I} & \text{for} \quad L_c = 0 \\ \exp\left(-\frac{\mathbf{2d}}{\lambda L_c}\right) & \text{for} \quad 0 < L_c < \infty \\ \mathbf{1} & \text{for} \quad L_c \to \infty \end{cases} \tag{12}$$

Source strength normalization

With an increasing number of sub-sources the SPL in the far-field generally increases (depending on their coherence, phase relationship, etc.). Since we want the total SPL to be independent of the number of sub-sources, we normalize their power. The normalization can be derived

by analytically calculating the CSM entries for the given sub-sources s_i with amplitude A_i and phase Q_i

$$s_i = A_i Q_i \,. \tag{13}$$

The coherence between the sub-sources is given by

$$Q_i Q_j = \eta_{ij} \tag{14}$$

with $0 \leq |\eta|^2 \leq 1$. Then we use for the n-th sensor the a-th Greens function g on the i-th sub-source s. The the microphone signal then is

$$p(x_n) = \sum_i g_{ia}(s_i) \,. \tag{15}$$

We rewrite the evaluation of the Greens function as a complex variable G_{ia}. Now, for illustration purposes we will calculate the Cross Spectral Density (CSD) for two sub-sources $s_{1,2}$, and two sensors $n_{a,b}$.

$$n_a = A_1 Q_1 G_{1a} + A_2 Q_2 G_{2a} \tag{16}$$

$$n_b = A_1 Q_1 G_{1b} + A_2 Q_2 G_{2b} \tag{17}$$

The CSD is defined as $C_{ab} = n_a n_b^*$.

$$C_{ab} = A_1^2 G_{1a} G_{1b}^* + A_1 A_2 G_{1a} G_{2b}^* \eta + A_1 A_2 G_{2a} G_{1b}^* \eta + A_2^2 G_{2a} G_{2b}^* \tag{18}$$

For this paper we assume that all sub-sources have the same amplitude $A_i = A$ and rewrite $G_{ij} G_{kl} = G_{ij,kl}^2$, thus

$$C_{ab} = A^2 G_{1a,1b}^2 + A^2 G_{1a,2b}^2 \eta + A^2 G_{2a,1b}^2 \eta + A^2 G_{2a,2b}^2 \tag{19}$$

$$= A^2 (G_{1a,1b}^2 + \eta (G_{1a,2b}^2 + G_{2a,1b}^2) + G_{2a,2b}^2) \,. \tag{20}$$

For the Power Spectral Density (PSD) C_{aa} this gives

$$C_{aa} = 2A^2 (1 + \eta G_{1a,2a}^2) \,. \tag{21}$$

Here we can see that the amplitude of the super positioned sources depends on their level of coherence. If the amplitude is supposed to be constant, it has to be normalized by $M^{\overline{\eta}}$. For a constant amplitude of A^2 for $M = 2$ sources this gives

$$A^2 = 2 \left(\frac{A}{2^{\overline{\eta}}} \right)^2 (1 + \eta) \tag{22}$$

This is the case for

$$\overline{\eta} = \frac{1}{2} \log_2 (2(\eta + 1)) \,. \tag{23}$$

Here we derived the normalization exponent $\overline{\eta}$, which explicitly depends on the coherence between the sub-sources. If C_{aa} is incoherent we simply expect $C_{aa} = 2A^2$ (see eq. 10), so the normalization exponent is $\overline{\eta} = 0.5$. For coherent sources we get a normalization exponent of

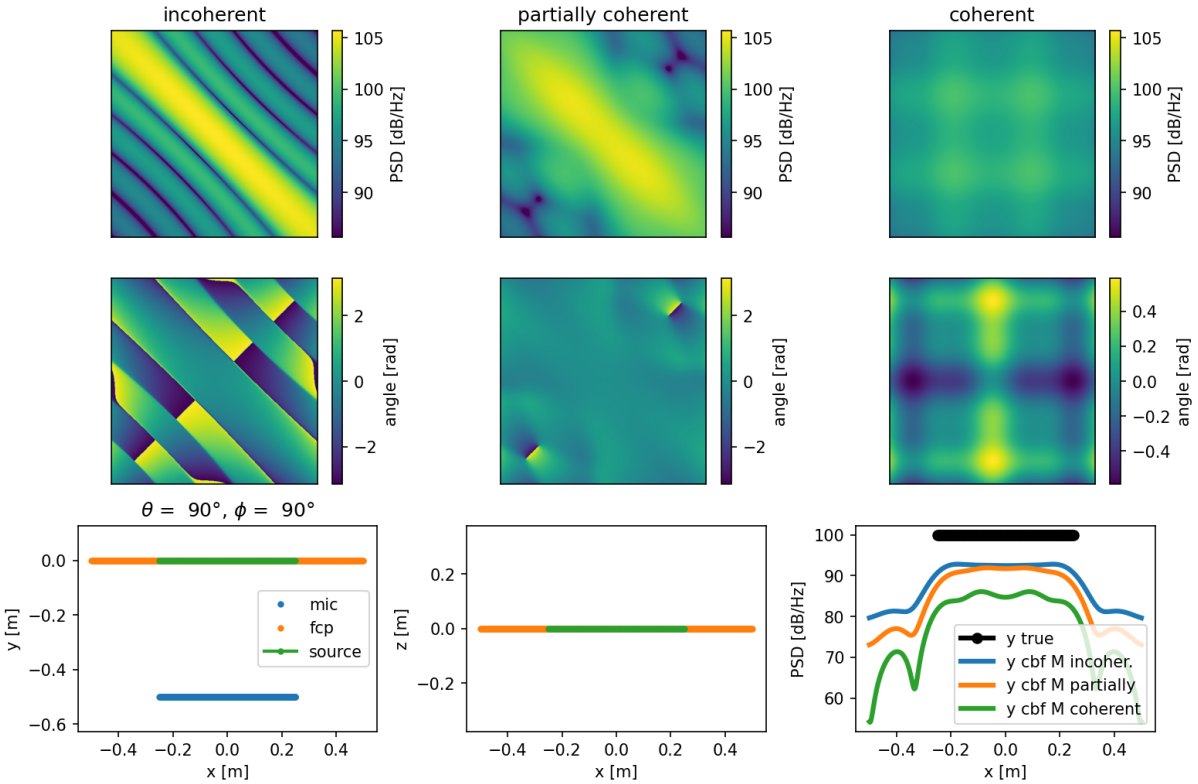

Figure 1: The Figure shows a comparison of an incoherent, a partially coherent, and totally coherent line-source with $l = 0.5\,\mathrm{m}$, $f = 4096\,\mathrm{Hz}$, $L_C = 3$, $PSD = 100\,\mathrm{dB}$. The top two rows display the coressponding CSMs, the bottom row shows the setup. The last image on the bottom row shows the CBF results based on these CSMs, assuming uncorrelated, incoherent sources.

$\overline{\eta} = 1$. It can easily be seen that this derivation works for M sources and results in a normalization of $1/M^{\overline{\eta}}$, thus $1/\sqrt{M}$ for incoherent and $1/M$ for coherent sub-sources. For a varying coherence level in Γ (e.g. with eq. 12) and under the condition that Γ is Hermetic we can simply use the averaged coherence matrix without its diagonal

$$\overline{\eta} = \langle |\gamma_{ij}^2| \rangle \quad \text{for} \quad i \neq j\,. \tag{24}$$

It follows for a distributed source with total amplitude A the normalized sub-source amplitude $\hat{s}_0 = A/2^{\overline{\eta}}$ in eq. 11. Figure 1 shows for a line source how the coherence length affects the CSM, and thus, the reconstructed SPL using conventional beamforming.

2.4 Synthetic measurement setup

For this paper we will use a simple 1.5D (x,y)-measurement setup, that consists of an equidistant $N = 5$ microphone setup $x = [-0.25\,\mathrm{m},\ldots,0.25\,\mathrm{m}]^T$, $y = -0.5\,\mathrm{m}, z = 0\,\mathrm{m}$ and 15 focus points $x = [-0.5\,\mathrm{m},\ldots,0.5\,\mathrm{m}]^T$, $y = 0\,\mathrm{m}, z = 0\,\mathrm{m}$. The measurement setup and the corresponding array

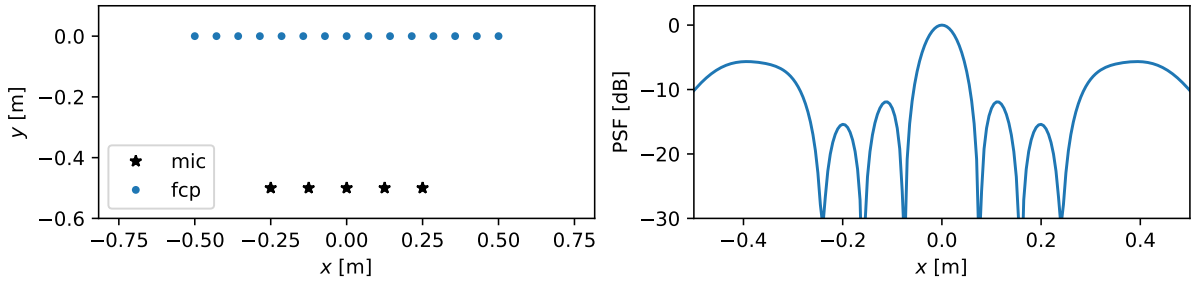

Figure 2: Standard measurement setup. Left: shows the array and focus points setup. Right: shows the corresponding PSF for $f = 4096\,\mathrm{Hz}$.

Point Spread Function (PSF) is displayed in Figure 2 for $f = 4096\,\mathrm{Hz}$.

3 ANN - Grid-based approach

For the grid-based approach, we will evaluate the performance of a few typical architectures. In particular, we want to explore general architectural design choices for future research.

3.1 Universal function approximator

An ANN can be seen as a universal function approximator, that learns a mapping implicitly from training data. The ANN will, if its learning capacity is large enough and the amount of training data approaches infinity, learn any function. For beamforming, the problem can be derived as follows. Given the propagation operator $\mathbf{T}$ of dimension microphones $N\times$ focus points M, the source vector q of dimension focus points, and the vectorized CSM c of dimension N^2 the forward problem is

$$\mathbf{T}q = c\,. \tag{25}$$

The propagation operator can be derived with $\mathbf{T} = \mathbf{H}^* \odot \mathbf{H}$, where $\odot$ is the Khatri-Rao product. Given that for the beamforming problem c is known, q is wanted, the problem can be solved by obtaining $\mathbf{T}^{-1}$

$$q = \mathbf{T}^{-1}c\,. \tag{26}$$

$\mathbf{T}$ has the rank $\min((N^2-N)/2, M)$. Thus, $\mathbf{T}$ is injective for $M \leq (N^2-N)/2$. However, $\mathbf{T}$ can be badly conditioned, see Figure 3, especially for the bijective case. In theory, the ANN should be able to learn the inverse Propagation operator while T^{-1} is injective. However, since the source vector q is often sparse, the ANN is expected to implicitly a regulation that might improve the performance on source vectors with few real sources.

3.2 Metrics

To evaluate the grid-based ANN results we define three metrics for sparse beamforming maps, which are inspired by Lehmann et al. [17], and Herold and Sarradj [11]. For sparse maps, in the sense that the number of non-zero Pa^2 entries is small compared to the number of grid points, the

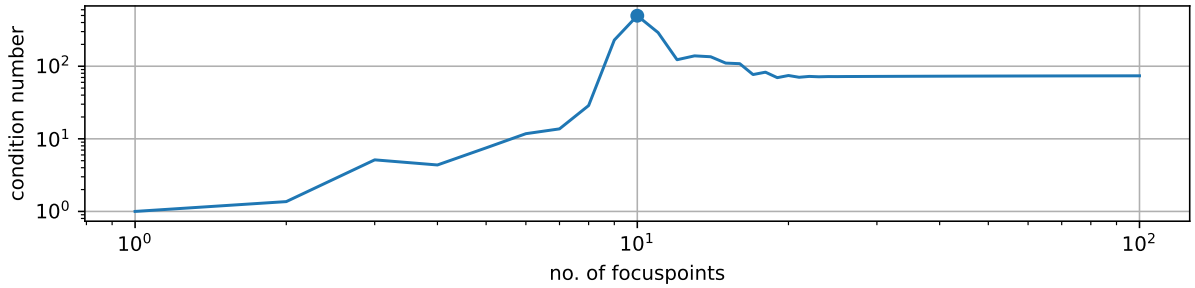

Figure 3: Conditioning number of the forward operator $\boldsymbol{T}$ for $N = 5$ microphones $x = [-0.25\,\mathrm{m}, \ldots, 0.25\,\mathrm{m}]^T$ and 15 focus points $x = [-1\,\mathrm{m}, \ldots, 1\,\mathrm{m}]^T, \Delta y = 0.5\,\mathrm{m}$, $f = 3400\,\mathrm{Hz}$. The case where $M = (N^2 - N)/2$ (T is bijective) is marked with a dot.

task contains classification, i.e. finding the correct focus points where there exists a source, and regression, i.e. finding the correct Sound Pressure Level (SPL). For the classification, we define a lower threshold (e.g. $L_T = 0\,\mathrm{dB}$), above which a focus point y is classified as a source S, and below which a focus point is classified as no source $\neg S$. Then we can define the sensitivity and specificity of the method given the number of True Positives (TP) $S_{\text{true}} \wedge S_{\text{pred.}}$, True Negatives (TN) $\neg S_{\text{true}} \wedge \neg S_{\text{pred.}}$, False Positives (FP) and False Negatives (FN).

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{27}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{28}$$

The sensitivity describes how many the real sources are identified, the specificity describes how many of the no sources are identified. Additionally, we need an error estimate for the regression task. We want the error to be relative (in decibel), and independent from the FP and FN classifications. Thus, we use the Source Strength Deviation (SSD) [17] as an Mean Absolute Error (MAE) between the estimated PSD and true PSD for the TP, and the average Noise to Signal Ratio (NSR) for the TN.

$$\text{SSD} = \left\langle \left| \text{PSD}_{\text{est.}}(y_{\text{TP}}) - \text{PSD}_{\text{true}}(y_{\text{TP}}) \right| \right\rangle \tag{29}$$

$$\text{NSR} = \frac{\left\langle \widehat{\text{PSD}}_{\text{est.}}(y_{\text{TN}}) \right\rangle}{\max\left(\widehat{\text{PSD}}_{\text{true}}(y_{S_{\text{true}}}) \right)} \tag{30}$$

We use the NSR instead of the typical Signal to Noise Ratio (SNR), because in the best case scenario the modified PSD (see section 3.3 and eq. 32) of the noise is zero, which we can only use as a numerator (the signal is non-zero by definition). The NSR captures how low the average PSD of an FP is, compared to the maximum PSD in the ground truth. We also want to know how close the estimated sources are able to reconstruct the given CSM. Thus, we solve the forward problem given the predicted sources to obtain a predicted CSM C_{pred} and define the

reconstruction error (RE).

$$\mathrm{RE} = \langle |C_{ij,\mathrm{true}} - C_{ij,\mathrm{pred.}}| \rangle \quad \text{for} \quad i > j \tag{31}$$

We use the MAE to keep it coherent with the SSD formulation. The RE provides an overall grasp of how well the combined classification and regression task is performed. Note, that the RE reconstruction error is not normally distributed, but Gamma distributed (not shown within this paper). Thus, with the Gamma function's parameter scale k and shape σ the mean error over all samples is $\langle \mathrm{RE} \rangle = k\sigma$ and its corresponding standard deviation is $\sqrt{k\sigma^2}$.

In summary, the RE gives a SPL-weighted single metric for both the classification and regression results. The sensitivity and specificity together describe the classification results. The SSD and NSR describe the regression results, based on the classification results.

3.3 ANN architecture

For the ANN architectural design, we will test different input and output configurations, described in the following.

CSM input dimensions

Given that the microphone array is structured so that every inner microphone has $3^D - 1$ neighbors in the D-dimensional array, we can reshape the CSM into an $(N_x \times N_y \times N_z)^2$ tensor. For a $1D$-array this is always the case. This allows for two options. First, the upper diagonal of the CSM is reshaped into a row. Second, the full CSM is used in tensor form.

Handling of complex input

Since the CSM is $C \in \mathbb{C}$, there are three options. Recently, complex-valued neuronal networks were established [4], which allow to directly input the complex CSM. Alternatively, the CSM can be split into their real and imaginary part. Or third, the CSM can be split into magnitude and phase. It should be noted, that only the first two options are holomorphic mappings. If the CSM input in tensor form is chosen in combination with a separation of the complex values, we can either add an additional dimension for the two real-valued CSM parts or use the upper diagonal for e.g. the real part and the lower diagonal for the imaginary part.

In- and output layer transformation

Beamforming results are typically evaluated in decibel. However, we cannot directly predict the resulting maps in decibel, due to the sparseness of the target map. Most of the target values will be $y_{\mathrm{true}} = 0\,\mathrm{Pa}^2 = -\infty\mathrm{dB}$. First, we will test a modified SPL formulation with

$$\widehat{\mathrm{SPL}} = 10\log_{10}\left(\frac{p^2}{p_0^2} + \varepsilon\right). \tag{32}$$

Figure 4 compares the normal and modified SPL definition for $\varepsilon = 1$ which will be used throughout this paper. This formulation allows for normal decibel operations (e.g., the addition of same

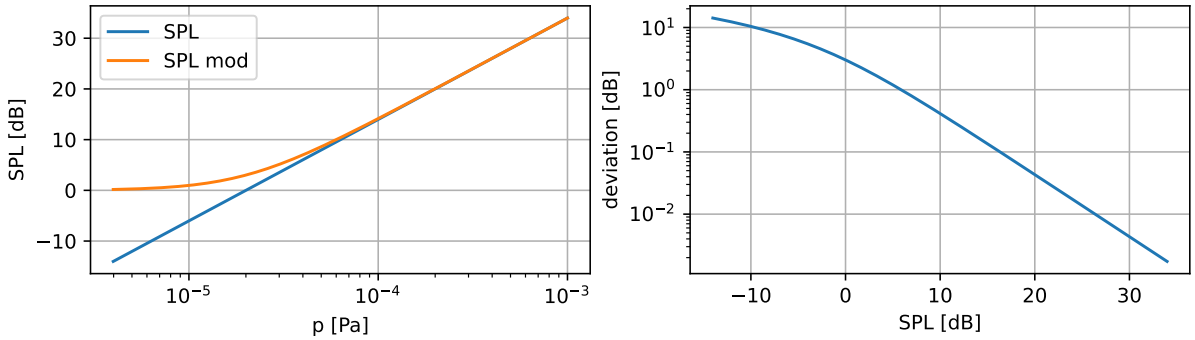

Figure 4: Left: Comparison of normal and modified SPL according to eq. 32. Right: Deviation between the two formulations.

levels should give +6 dB) with less than 1 % deviation above around $L > 25\,\mathrm{dB}$, and fixes the infinity issues. If we are interested in the correct detection of lower SPL, we can always adjust ε towards lower values. Second, we will test a natural logarithm with

$$p^2_{\log} = \log\left(p^2 + 1\right) . \tag{33}$$

Third, we will test to predict y_{true} directly in Pa^2. The corresponding input CSM will be transformed according to the output layer. Additionally, the input and outputlayer are then normalized to $0 \leq |y| \leq 1$, $0 \leq |C| \leq 1$.

Loss function

A typical loss function for the regression is the Mean Squared Error.

$$\mathrm{MSE} = \langle |y_{\mathrm{true}} - y_{\mathrm{pred.}}|^2 \rangle \tag{34}$$

However, since the output vector is very sparse, we can add a weighting function w to prefer the correct SPL at focus points that feature a true source

$$\mathrm{wMSE} = \langle w|y_{\mathrm{true}} - y_{\mathrm{pred.}}|^2 \rangle \tag{35}$$

For the total number of focus points that a true source N_S and the total number of focus points that do not have a source $N_{\neg S}$ the weight then is

$$\mathrm{w} = \begin{cases} 1/N_S & \text{for} \quad y_{\mathrm{true}} \in S \\ 1/N_{\neg S} & \text{for} \quad y_{\mathrm{true}} \in \neg S \end{cases} . \tag{36}$$

Optionally, to prevent the sparse output vector, we can use a Fourier Transformation to predict the SPL as a wavenumber. According to the handling of the complex values of the CSM, we can either use a fully complex ANN, or split the complex numbers in e.g. real and imaginary part. When adding an additional dimension for a CNN, this allows to use convolutions separately on the real and imaginary part throughout the network.

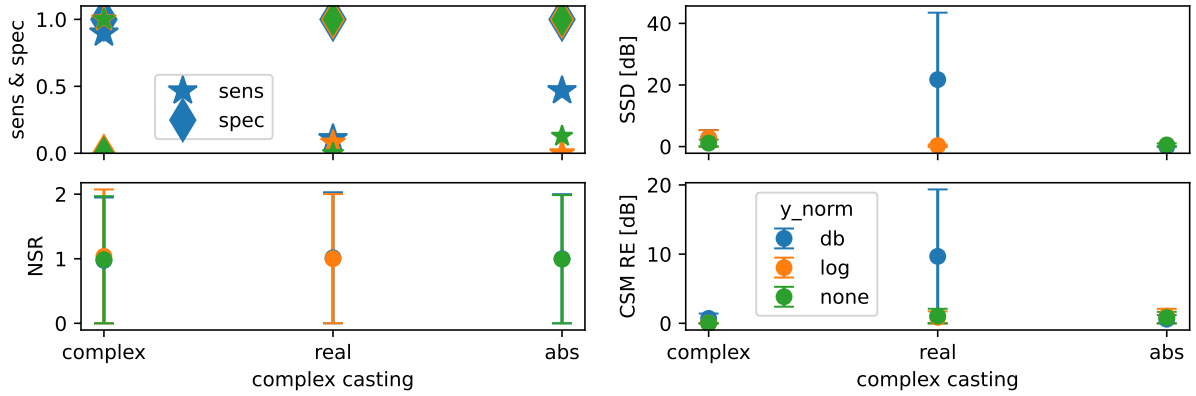

Figure 5: Error metrics of section 3.2 for the baseline model with variation of the complex input handling (see sec. 3.3), and the input and output transformation (see sec. 3.3) for $N_S = 1$ source.

Multipoles

If we want to predict multipoles we have two options. We either predict the SPL on separate focus grids for each pole or we use one focus grid for the SPL and an additional grid on which the pole order is predicted. Only the first option can handle different poles on the same location. However, this results in $(3 \times M)$ extra variables (SPL, and two rotations of the multipole) for each multipole beyond the monopole. As we will discover in the next sections, the grid-based problem is ill-posed. Thus, we will only focus on compact monopoles for the grid-based approach.

4 ANN architectures

Since these architectural design possibilities provide endless combination options, we will start with a simple baseline model and test several parameters. Then, based on the observations we will define and test multiple architectures. All architectures are trained on 10^6 training samples for 100 epochs with a $0.9/0.1$ validation split, and the error metrics are evaluated on 10^3 test samples. Note, that for the grid-based approach, sources are only located on the focus points to prevent aliasing and assignment ambiguities.

4.1 Baseline Model

The baseline model is a simple fully connected (FC) network, with Rectified Linear Activation Functions (ReLu), three Hidden Layers (HL), and 256 Neurons Per Layer (NPL) with a MSE loss. We use $N = 5$ microphones and $M = 15$ focus points. On each grid, there is only $N_S = 1$ real source with $50\,\text{dB} \leq \text{SPL} \leq 100\,\text{dB}$ and the threshold for a source $L_T = 30\,\text{dB}$. Note, that the complex ANN has around twice the number of learnable parameters (300k), compared to real-valued ANNs (150k).

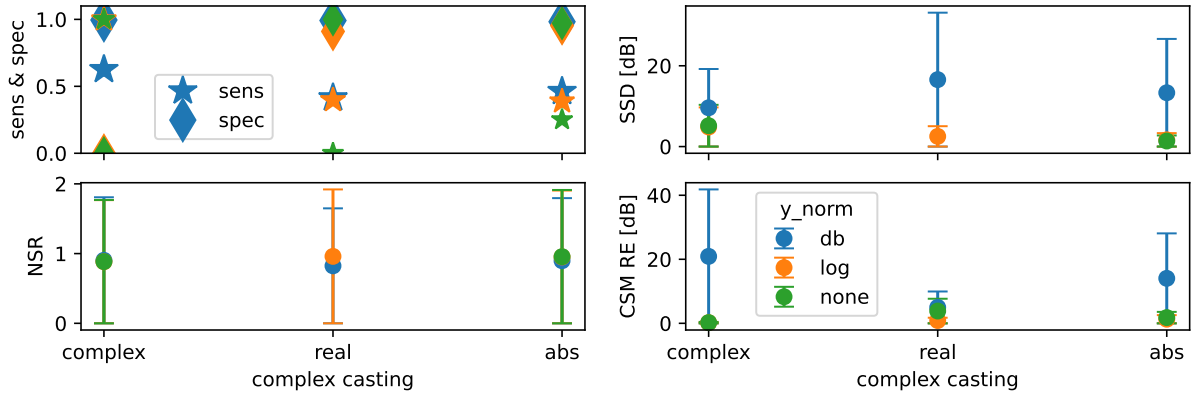

Figure 6: Error metrics of section 3.2 for the baseline model with variation of the complex input handling (see sec. 3.3), and the input and output transformation (see sec. 3.3) for $N_S = 5$ sources.

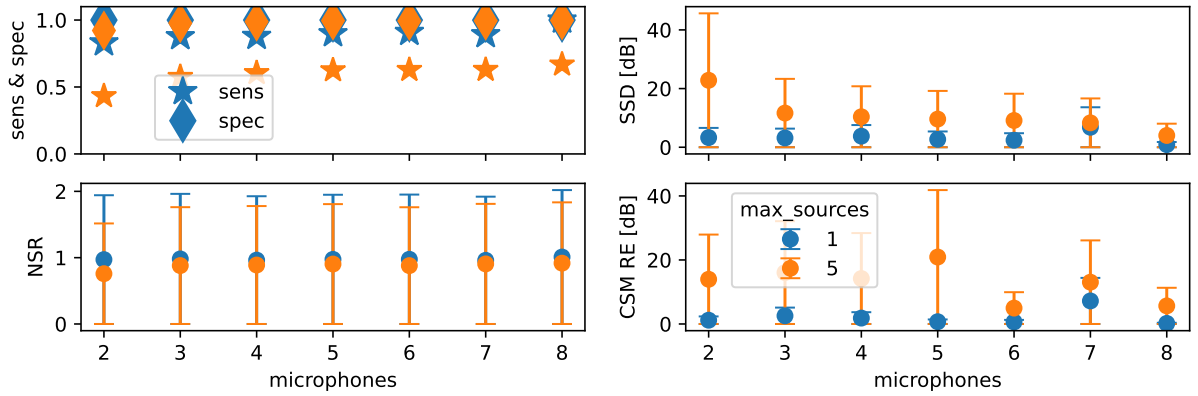

Figure 7: Error metrics of section 3.2 for the complex baseline model with dB transformation and an increasing number of microphones.

For the compared configurations, the complex ANN with the modified dB-transformation according to eq 32 outperforms the other configurations. The combination of a low CSM RE and a NSR close to unity means that in some instances the ANN predicted the correct source strength on a wrong, but close to the true focus point. Figure 6 shows the same setups for $N_S = 5$ sources. Again, the complex ANN in combination with the dB transformation outperforms the other models on the classification task. On the regression task, the other transformation methods are more precise. The low CSM RE on the other models combined with the low classification accuracy shows, that the models are not able to produce sparse results. E.g., the complex ANN with no transformation predicts a source (that is $p^2 \geq 0\,\mathrm{Pa}^2$) on each focus point, but overall, the predicted pressure distribution is quite reasonable. Since we are both interested in the regression and classification, we will focus on the complex ANN in combination with the dB transformation in the following.

In sec. 3.1 we discussed the uniqueness of the solution, given the number of microphones

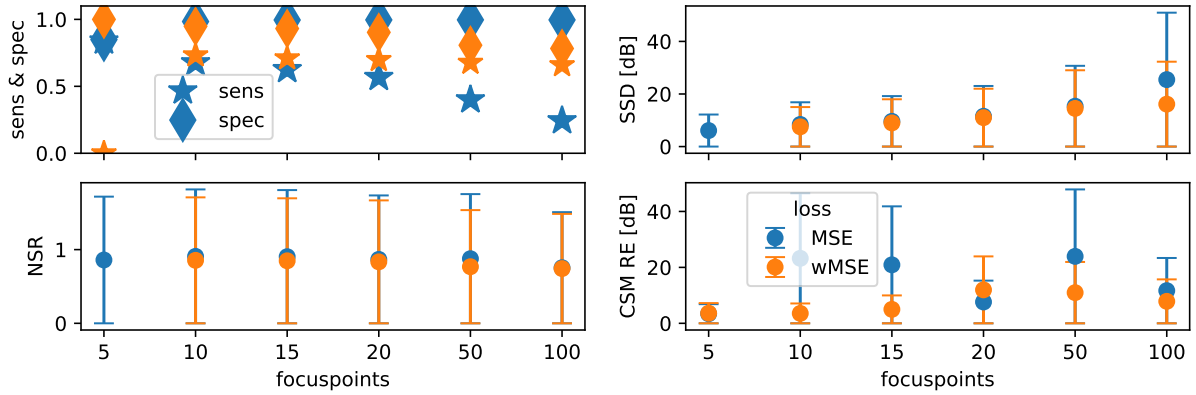

Figure 8: Error metrics of section 3.2 for the complex baseline model with dB transformation, the MSE and wMSE at an increasing number of focus points.

and focus points. Figure 7 shows the corresponding results for an increasing number of microphones. Given the $M = 15$ focus points, the problem has a unique solution for $N \geq 7$ microphones. For $N_S = 1$ the ANN achieves nearly perfect results at $N = 8$. Interestingly, for $N_S = 5$ the results are much worse. At $N = 7$ (for the approximately bijective case) we can see an increased CSM RE, which might be connected to the bad conditioning of the propagation matrix (see Figure 3). While the ANN does not achieve good results for the injective setups, the results for the underdetermined cases with $N < 7$ are impressive (even for $N = 2$ and $N_S = 5$, where the ANN receives only a single complex input, the ANN correctly classifies around 50 % of the sources).

Finally, we will evaluate the loss functions for the base model, see sec. 3.3. Figure 8 shows the results for the MSE and wMSE for an increasing number of focus points at $N = 5$, $N_S = 5$.
The wMSE outperforms the regular MSE in all cases (except for $N = 5$ where all focus points are weighted with zero for $N_S = 5$).

In summary, the combination of a wMSE for the loss function, a complex-valued ANN with dB transformation give overall the most promising results. Based on this observation we will define different ANN architectures in the following.

4.2 ANN architectures

In this section, we will compare different ANN architectures. The architectures will be separated into the CSM input, the ANN encoder part, the ANN decoder part, and the output. For the input we will use the CSM row structure and tensor structure, for the encoder we will use fully connected (FC) and convolutional layers (CL), for the decoder we will use FC and CL, and for the output, we will use the focus points (FP), and the source wavenumber vector (WN) (with a MSE loss). The number of hidden layers of the encoder (EHL) and decoder (DHL) is stated, for a FC network (which does not really have an encoder-decoder structure), the number of HL is simply listed in EHL, see Table 2. Since the number of input and output neurons differ based on the given architecture, the numbers of hidden convolutional layers with a (3)-kernel and a

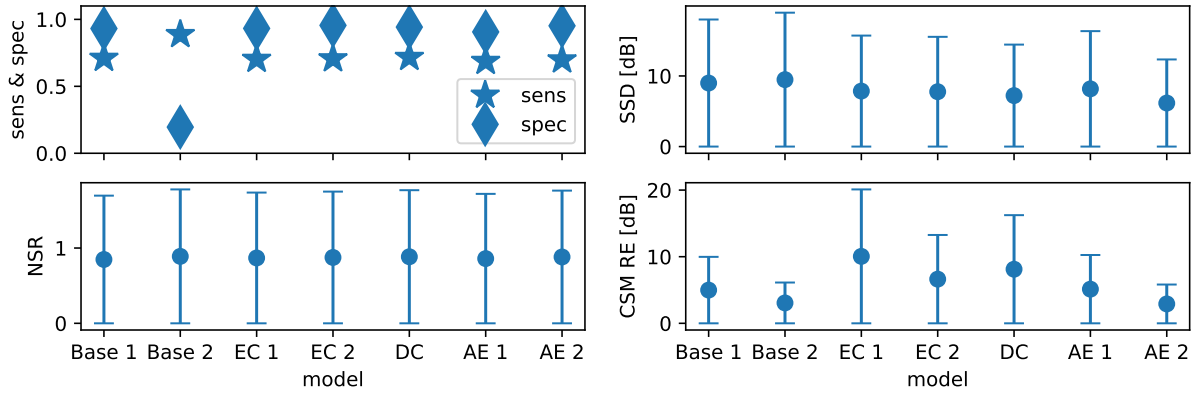

Figure 9: Error metrics of section 3.2 for the different model architectures from Table 2.

(1)-stride also differ. For the tensor input the CSM is 2D, and the convolution kernel is (3×3) with a (1×1)-stride.

model	CSM shape	encoder	EHL	decoder	DHL	output	loss
Base 1	row	FC	3	FC		FP	wMSE
Base 2	row	FC	3	FC		WN	MSE
EC 1	row	CL	3	FC	3	FP	wMSE
EC 2	tensor	CL	2	FC	3	FP	wMSE
DC	row	FC	3	CL	5	FP	wMSE
AE 1	row	CL	3	CL	5	FP	wMSE
AE 2	tensor	CL	2	CL	7	FP	wMSE

Table 2: ANN architectures for benchmark.

Figure 9 shows the resulting metrics for the model architectures from Table 2. Except for the Baseline model that predicts the complex wavenumber vector all models perform similarly. The Auto Encoder 2 model has a slight advantage in terms of SSD and CSM RE, which makes it the best model overall. Figure 10 shows three example configurations from the test set and the AE 2 models predictions.

4.3 Final words on the grid-based approach

As we have seen the results on grids are rather mediocre. The reason for this might be, that the ANN is trying to learn the inverse of the propagation operator. However, the community of inverse problems pointed out, that this is a sub-optimal task for ANNs [3], especially if we have in fact information about the operator (e.g., monopole assumption). It also suffers from the difficult mapping from distributed sources on focus points, as well as spatial aliasing. Furthermore, the result is a high dimensional map, in which real sources and their spectra have to be identified [10].

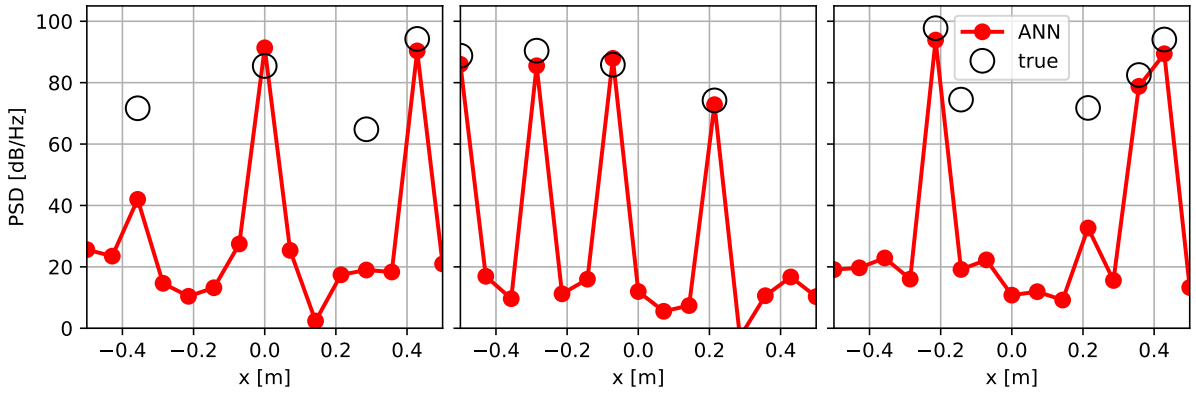

Figure 10: Three example results of the test set for the AE 2 model.

5 ANN - Grid-free approach

Instead, we will focus now on grid-free ANN architectures. The main advantage is, that we can directly predict a source object. A source object can have multiple properties, such as its spectrum, location, directivity, spatial compactness, and corresponding coherence length. We will treat those source objects as members of a set, and each member contains a vector with the mentioned variables. This means, that the source objects do not have a particular order (permutation invariant), but the contained vectors do (not permutation invariant). This problem has been assessed recently in machine learning with deep sets [27], and was particularly successful in object detection and classification tasks [6, 14, 18] in combination with transformer networks [25]. The key to this problem is a permutation invariant loss function, which implicitly or explicitly matches the predicted set with the target set before calculating the loss. A precise matching can be achieved with the Hungarian Algorithm, and a lower error bound can be achieved with the Chamfer Loss. In the following, we will explore how a grid-free network can be realized. However, the following part is a work in progress, and no final ANN architecture is presented, which incorporates all of the described methods.

For the first shot at this problem we will again only regard a single monopole source with arbitrary location $0.1\,\mathrm{m} \geq r \geq 10\,\mathrm{m}$, $-\pi/2 \geq \theta \geq \pi/2$ and $50\,\mathrm{dB} \geq \mathrm{SPL} \geq 100\,\mathrm{dB}$. Since we are not restricted by the focus points size anymore, we will use a 2D source distribution (in (x,z), thus, with depth information) for the 1D array (in x). Thus, the ANN has three output neurons: The SPL, the x-coordinate, and the z-coordinate. However, since we choose a wide range of possible radii, we will transform the coordinates into spherical coordinates and predict the angle θ and radius r (logarithmically). This is a regression task and we can calculate an error for all of these parameters independently. To equally weight all errors in the loss function, the variables are independently normalized to $[0,\ldots,1]$, and then the average MSE is calculated.

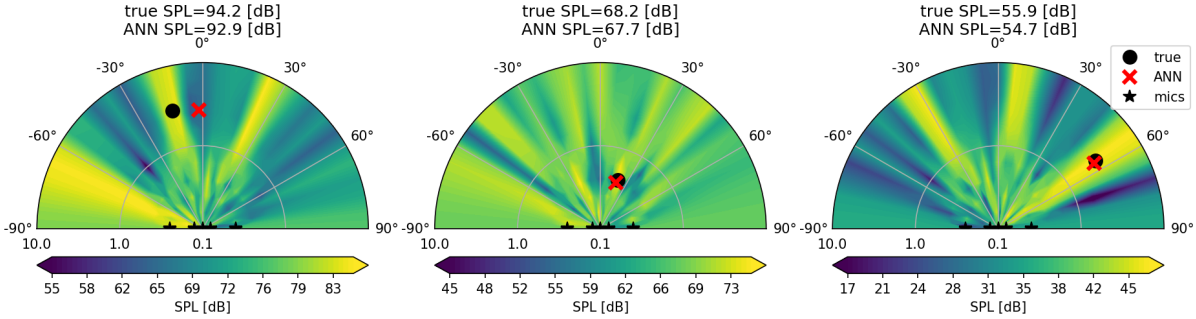

Figure 11: Three example results of the test set for grid-free approach and three normalized output neurons: SPL, $\log_{10}(r)$, and θ. The underlying color shows how conventional beamforming (BF) with steering vector formulation I [23] (no distance weight) performs for the given configuration.

5.1 Grid-free error metrics

For the Grid-Free (GF) single source regression problem the error metrics are straight forward. The Source Strength Deviation is

$$\mathrm{SSD}_{\mathrm{GF}} = \left\langle \left| \mathrm{PSD}_{\mathrm{pred.}} - \mathrm{PSD}_{\mathrm{true}} \right| \right\rangle . \tag{37}$$

The Angular Deviation (AD) is

$$\mathrm{AD}_{\mathrm{GF}} = \left\langle \left| \theta_{\mathrm{pred.}} - \theta_{\mathrm{true}} \right| \right\rangle . \tag{38}$$

Since we using a source distribution over a large range of radii, we will use the Relative Radius Error, based on the true radius.

$$\mathrm{RRE}_{\mathrm{GF}} = \left\langle \frac{\left| r_{\mathrm{pred.}} - r_{\mathrm{true}} \right|}{r_{\mathrm{true}}} \right\rangle . \tag{39}$$

5.2 Single source

Figure 11 shows three examples from the test set for a FC complex model, with 3 HL, and 256 NPL (like the base model in the grid-based approach). Thus, this is a regression model only. As we can see, the results are much more promising than the single source configurations from the grid-based model. Table 3 gives the average test set errors for the model. Additionally, the model was trained on dipoles with $-\pi/2 \geq \theta_D \geq \pi/2$, $\phi_D = 0$. Note, that the dipole angle is not uniquely determinable from the CSM (it is π-symmetric), and it might make sense to only predict the absolute angle for one source.

5.3 Multiple sources

Given we want to predict multiple sources, we have to define a maximum number of possible source objects O. Each object has multiple properties, including the existence (zero for no

	$\mathrm{SSD}_{\mathrm{GF}}[\mathrm{dB}]$	$\mathrm{RRE}_{\mathrm{GF}}[\%]$	$\mathrm{AD}_{\mathrm{GF}}(\theta)[^\circ]$	$\mathrm{AD}_{\mathrm{GF}}(\theta_D)[^\circ]$
monopole	0.94 ± 0.96	9.62 ± 9.88	3.39 ± 5.71	-
dipole	2.61 ± 3.62	22.66 ± 33.46	11.56 ± 15.19	30.74 ± 39.74
$\leq$5 monopoles	7.53 ± 10.04	89.21 ± 157.42	32.47 ± 41.29	-

Table 3: Errors of the grid-free models. For the dipole model, there is the additional dipole orientation angle θ_D. The last row are the errors from the permutation invariant ANN with up to five monopole sources.

source, and one for a source), the radius r, the θ angle, and the SPL. Since unlike on a fixed grid it is not clear for an ANN to assign which output neurons to which source object, we need a permutation invariant output. We can achieve this by calculating all $O!$ permutations of the source objects. Then we calculate the loss of these permutations independently and choose the minimum loss for the best match with the true set. The loss for a sample of dimension $[O!, O, 4]$ is

$$\mathrm{loss} = \min_{O!} \left(|\mathrm{CE}| + \langle \mathrm{MSE}(r, \theta, \mathrm{SPL}) \rangle \right) , \tag{40}$$

where MSE is the regression variable averaged Mean Squared Error and CE is the Binary Cross Entropy of the classification variable. Since it makes no sense to calculate a regression error on the source objects that are classified as no source, we mask the corresponding variables. We set each regression variable in $y_{\mathrm{true}}(\neg S_{\mathrm{true}})$ and $y_{\mathrm{pred.}}(\neg S_{pred.})$ to zero.

For a first shot at a permutation invariant ANN we will use a FC complex Neuronal Network, with NPL=256, HL=3, and $O = 5$ (there are up to five sources, the actual number is equally distributed between 1 and 5 per sample). We use two complex HL with Cartesian ReLu, then a magnitude (absolute value) layer, and then a real-valued FC layer without an activation function, because the output needs to be able to predict negative values $]-\infty, \infty[$ for the Binary Cross Entropy. According to eq. 27 and 28 the test samples sensitivity$= 0.80$ and the specificity$= 0.81$. Table 3 gives the average regression errors. Figure 12 shows a sample from the test set. We can see that the permutation matching works well and the classification task is performed with reasonable accuracy, especially for the large SNR=50 dB. The regression is not performed well for the given setup. The reason for this might be the very simple ANN design and the mismatch between necessary activation functions for the different tasks of the output neurons.

6 Summary and outlook

In this paper, we investigated the generation of synthetic sources in the frequency domain. We showed how to generate any source distribution with any level of coherence between the sources. In particular, we covered the generation of monopoles and dipoles. For future work, it would be very useful to include the effects of a real-world measurement in the CSM, such as the Welch error based on the number of averages, microphone self-noise, and wind tunnel noise and use real data in the test set. We then covered the topic of grid-based ANN beamformers that predict the source distribution on a fixed focus grid. First, we discussed several metrics that are enable an interpretation of the sparse results. Then, we evaluated multiple ANN architectures.

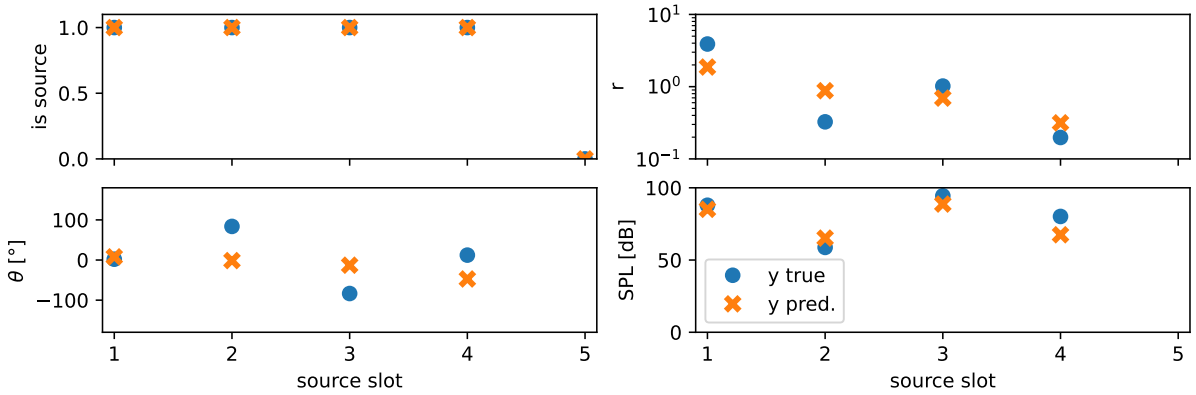

Figure 12: Example result of the test set for grid-free approach with permutation matching. There are $O = 5$ possible source slots with four variables each.

We saw that several issues, such as the increasing sparseness on an increasing focus grid and the handling of the complex CSM, can be handled using a modified SPL formula, complex-valued ANNs, and weighted losses. However, the grid-based architecture has conceptual flaws, as it is difficult to handle multiple source types at once, suffers from aliasing, and does not scale well for real-world focus grid sizes. Also, we explored that the forward operator, of which the ANN is supposed to learn the inverse, is badly conditioned. This can result in large CSM reconstruction errors. We then explored how we can predict grid-free source objects. This is known as a tensor-to-set problem and uses a two-step loss function. First, a set member matching is performed, since a set is permutation invariant. Then, the loss for the best match is calculated. While this topic is still a work in progress, preliminary results showed that this method has potential. It also fixes the inherent problems of the grid-based approach, such as the discrete mapping of distributed sources, and the detection of real sources from the resulting beamforming maps. For future work it would be useful to incorporate advanced ANN designs to improve the regression results and perform the beamforming for multiple frequencies at once.

References

[1] T. Adali, P. J. Schreier, and L. L. Scharf. "Complex-valued signal processing: The proper way to deal with impropriety." *IEEE Transactions on Signal Processing*, 59(11), 5101–5125, 2011. doi:10.1109/tsp.2011.2162954.

[2] J. Antoni. "A bayesian approach to sound source reconstruction: Optimal basis, regularization, and focusing." *The Journal of the Acoustical Society of America*, 131(4), 2873–2890, 2012. doi:10.1121/1.3685484.

[3] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. "Solving inverse problems using data-driven models." *Acta Numerica*, 28, 1–174, 2019. doi:10.1017/s0962492919000059.

[4] J. A. Barrachina. "Complex-valued neural networks (cvnn).", 2021. doi:10.5281/zenodo.4452131.

[5] T. F. Brooks and W. M. Humphreys. "A deconvolution approach for the mapping of acoustic sources (DAMAS) determined from phased microphone arrays." *Journal of Sound and Vibration*, 294(4-5), 856–879, 2006. doi:10.1016/j.jsv.2005.12.046.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. "End-to-end object detection with transformers.", 2020. doi:10.48550/ARXIV.2005.12872.

[7] 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." *Applied Acoustics*, 177, 107947, 2021. doi:10.1016/j.apacoust.2021.107947.

[8] B. Dong, J. Antoni, A. Pereira, and W. Kellermann. "Blind separation of incoherent and spatially disjoint sound sources." *Journal of Sound and Vibration*, 383, 414–445, 2016. ISSN 0022-460X. doi:https://doi.org/10.1016/j.jsv.2016.07.018.

[9] B. Dong, J. Antoni, and E. Zhang. "Blind separation of sound sources from the principle of least spatial entropy." *Journal of Sound and Vibration*, 333(9), 2643–2668, 2014. ISSN 0022-460X. doi:https://doi.org/10.1016/j.jsv.2013.12.011.

[10] A. Goudarzi, C. Spehr, and S. Herbold. "Automatic source localization and spectra generation from sparse beamforming maps." *The Journal of the Acoustical Society of America*, 150(3), 1866–1882, 2021. doi:10.1121/10.0005885.

[11] G. Herold and E. Sarradj. "Performance analysis of microphone array methods." *Journal of Sound and Vibration*, 401, 152–168, 2017. doi:10.1016/j.jsv.2017.04.030.

[12] T. Hohage, H.-G. Raumer, and C. Spehr. "Uniqueness of an inverse source problem in experimental aeroacoustics." *Inverse Problems*, 36(7), 075012, 2020. doi:10.1088/1361-6420/ab8484.

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

[14] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teh. "Set transformer: A framework for attention-based permutation-invariant neural networks." In *Internation Conference on Machine Learning*. arXiv, 2018. doi:10.48550/ARXIV.1810.00825.

[15] S. Y. Lee, J. Chang, and S. Lee. "Deep learning-based method for multiple sound source localization with high resolution and accuracy." *Mechanical Systems and Signal Processing*, 161, 107959, 2021. doi:10.1016/j.ymssp.2021.107959.

[16] S. Y. Lee, J. Chang, and S. Lee. "Deep learning-enabled high-resolution and fast sound source localization in spherical microphone array system." *IEEE Transactions on Instrumentation and Measurement*, 71, 1–12, 2022. doi:10.1109/tim.2022.3161693.

[17] M. Lehmann, D. Ernst, M. Schneider, C. Spehr, and M. Lummer. "Beamforming for measurements under disturbed propagation conditions using numerically calculated green's functions." *Journal of Sound and Vibration*, 520, 116638, 2022. doi:10.1016/j.jsv.2021.116638.

[18] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf. "Object-centric learning with slot attention.", 2020. doi: 10.48550/ARXIV.2006.15055.

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

[20] 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.

[21] W. G. Pinto, M. Bauerheim, and H. Parisot-Dupuis. "Deconvoluting acoustic beamforming maps with a deep neural network." *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, 263(1), 5397–5408, 2021. doi:10.3397/in-2021-3084.

[22] H.-G. Raumer, C. Spehr, T. Hohage, and D. Ernst. "Weighted data spaces for correlation-based array imaging in experimental aeroacoustics." *Journal of Sound and Vibration*, 494, 115878, 2021. doi:10.1016/j.jsv.2020.115878.

[23] E. Sarradj. "Three-dimensional acoustic source mapping with different beamforming steering vector formulations." *Advances in Acoustics and Vibration*, pages 1–12, 2012. doi:10.1155/2012/292695.

[24] P. Sijtsma. "Clean based on spatial source coherence. international journal of aeroacoustics." *International Journal of Aeroacoustics*, 6, 357–374, 2007. doi:10.1260/147547207783359459.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention is all you need.", 2017. doi:10.48550/ARXIV.1706.03762.

[26] P. Xu, E. J. G. Arcondoulis, and Y. Liu. "Deep neural network models for acousticsource localization." In *Berlin Beamforming Conference*. 2021.

[27] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola. "Deep sets.", 2017. doi:10.48550/ARXIV.1703.06114.

[28] P. Zavala, W. D. Roeck, K. Janssens, J. Arruda, P. Sas, and W. Desmet. "Generalized inverse beamforming with optimized regularization strategy." *Mechanical Systems and Signal Processing*, 25(3), 928–939, 2011. doi:10.1016/j.ymssp.2010.09.012.

[29] E. Zhang, J. Antoni, B. Dong, and H. Snoussi. "Bayesian space-frequency separation of wide-band sound sources by a hierarchical approach." *The Journal of the Acoustical Society of America*, 132(5), 3240–3250, 2012. doi:10.1121/1.4754530.