BeBeC-2024-S01

TOWARDS GENERALIZATION OF FREQUENCY-INVARIANT DEEP LEARNING MODELS FOR GRID-FREE SOURCE CHARACTERIZATION

Adam Kujawski and Ennes Sarradj
TU Berlin, FG Technische Akustik
Einsteinufer 25, 10587 Berlin, Germany

Abstract

Within recent years, several data-driven microphone array methods showed promising potential regarding their performance in accurately characterizing multiple sound sources from microphone array data. These methods have one thing in common: they were trained using virtually supervised learning. While excellent performance is frequently reported on synthetic data for virtually trained models, a typical observation in experimental applications is performance degradation due to the differently distributed data in the experimental domain. To date, the experimental generalization behavior of these methods has yet to be explored. Another largely unexplored aspect is the performance of grid-free data-driven methods when training with microphone array data from multiple frequencies using a single model architecture.

This work analyzes the characterization performance of a grid-free deep learning method that is trained with microphone array data from multiple frequencies and compares it to the performance of single-frequency trained models. Furthermore, the generalization behavior for the frequency-invariant method is examined in the virtual and experimental domains. A sizeable dataset is employed to obtain statistically meaningful results. The experimental data is based on the MIRACLE dataset, a recently published database containing measured impulse responses from a loudspeaker at various locations under anechoic conditions.

1. INTRODUCTION

Numerous data-driven microphone array methods for source mapping and characterization have recently been developed [8, 10], and the literature suggests that data-driven methods can become a valuable complement to the existing model-based microphone array methods [22]. One way to classify the various data-driven approaches is to divide them into hybrid and entirely data-driven methods. Hybrid methods support conventional microphone array methods through data-driven pre- or post-processing [5, 6, 19], or replicate microphone array methods to achieve faster solutions [3, 11], including the deconvolution of beamforming source mappings [4, 16, 25, 29]. Hybrid methods usually incorporate physical knowledge about the sound propagation model, either in the model itself or during data pre- or post-processing. Entirely data-driven methods do not require prior physical knowledge, predict the source characteristics based on the raw microphone array data, and learn relevant features for source characterization from the training data [2, 15, 21, 28]. Data-driven methods are predominantly trained using a virtually supervised learning [26, 27]. In supervised learning, the goal is to learn how to map the input data and available labels, i.e., the microphone array data and the source characteristics. Typically, training data is generated using acoustic simulations since ground-truth source characteristics for experimental measurements and sizable experimental datasets are difficult to obtain. When a virtually trained model is applied to experimentally obtained microphone array data, the model has to generalize to the (unknown) experimental domain. The shift between the virtual and the experimental domain will affect the model's performance, a common problem in machine learning applications [23, 30]. There exist various reasons causing the domain shift, including uncertainties in the measurement setup, the presence of noise, and the influence of the environment.

A coherent and promising observation across the available literature is that data-driven microphone array methods lead to faster and sometimes more accurate source mappings than conventional methods. Some methods exhibit superior performance at low frequencies, where many model-based methods become inaccurate [4, 15, 28, 29]. However, the promising results must be taken cautiously since many studies evaluated the characterization performance solely in the virtual domain [16, 21, 25, 29]. Until recently, no large-scale experimental datasets were available to evaluate the experimental generalization performance of data-driven methods. Studies using experimental data for evaluation usually rely on a single or a few measured test cases with a limited number of sound sources [2, 15]. Among the cited works, only [1, 18] employed experimental microphone array data for model training. However, they did not explicitly investigate the domain shift, consequently evaluating generalization solely within the same domain. Thus, the impact of the domain shift on the performance of deep learning models for source characterization is mainly unexplored but is a crucial aspect for practical applicability.

The Transformer model for grid-free source mapping [15] is used in this work to investigate the domain generalization performance. In the original publication [15], the model was trained using cross-spectral matrix (CSM) data from a single frequency or third-octave band. Applying the model to other frequencies requires retraining, which is time-consuming and often impractical. Some grid-based deep learning models operating on source maps obtained with Conventional Beamforming (CB) have demonstrated that a single model architecture can also be trained blindly with data from multiple frequencies, making them frequency-invariant [13, 19]. However, the beamforming map already contains pre-processed spatial information and provides a frequency-encoded representation due to the influence of the frequency-dependent microphone

array transfer function. Whether frequency invariance can be achieved for grid-free methods that directly operate on raw microphone array data is an open question.

The aim of this work is two-fold. Firstly, frequency-invariant training of the Transformer architecture is investigated. Therefore, the model is blindly trained with cross-spectral matrices (CSMs) of different frequencies, and its performance is compared to that of single-frequency trained models. This model is then used to examine experimental generalization in two ways. Firstly, the generalization performance of the virtually trained model is examined in the virtual and experimental domain under anechoic conditions. Secondly, the generalization performance is examined with respect to environmental mismatch, i.e., the virtually trained model is tested on experimental data containing specular reflections. A sizeable dataset is employed to obtain statistically meaningful results in the experimental domain, based on the Microphone Array Impulse Response Dataset for Acoustic Learning (MIRACLE), a recently published database[1] containing measured room impulse responses (RIRs) from a loudspeaker at various locations under anechoic conditions [12]. The source code of the Transformer model and the utilized datasets are openly available[2].

2. METHODOLOGY

The source characterization problem for noise sources with stochastic signals can be defined as follows. Given data from M spatially distributed microphones and J sound sources, the goal is to obtain an estimate for the ground-truth source characteristics $\mathscr{G} = \{s_j \mid j = 1,\dots,J\}$. In this work, $s_j \in \mathbb{R}^2 \times \mathbb{R}^+$ is represented by a tuple containing the source position $x_j \in \mathbb{R}^2$ of the j-th source in a planar observation area and the expectation value of the squared sound pressure

$$a(x_j,\omega) = E\left[p(x_j,x_0,\omega)p(x_j,x_0,\omega)^*\right] \tag{1}$$

at the reference position x_0. In this work, ω denotes the angular frequency, x_0 lies at the microphone location closest to the center of the microphone array, and $a(x_j,\omega)$ is referred to as the source strength.

Grid-based methods solve the source characterization problem by discretizing the observation area into I grid positions and calculating the source strength for each grid point. Grid-free methods, on the other hand, do not require spatial discretization of the observation area. In either way, a set of estimated source characteristics $\mathscr{E} = \{s_i \mid i = 1,\dots,I\}$ is obtained. While the number of estimates I usually differs from the number of sources J, assigning the estimates to the unknown ground-truth sources is necessary. In acoustic-testing applications, there is often knowledge regarding the potential regions at which sound sources can be expected. This information defines a region of interest (ROI) over which the squared sound pressure can be summed to obtain an estimate for the j-th ground-truth source strength

$$\hat{a}(x_j,\omega) = \sum_{x_i \in \mathrm{ROI}} \hat{a}(x_i,\omega). \tag{2}$$

[1] https://doi.org/10.14279/depositonce-20106
[2] https://github.com/adku1173/BeBeC2024

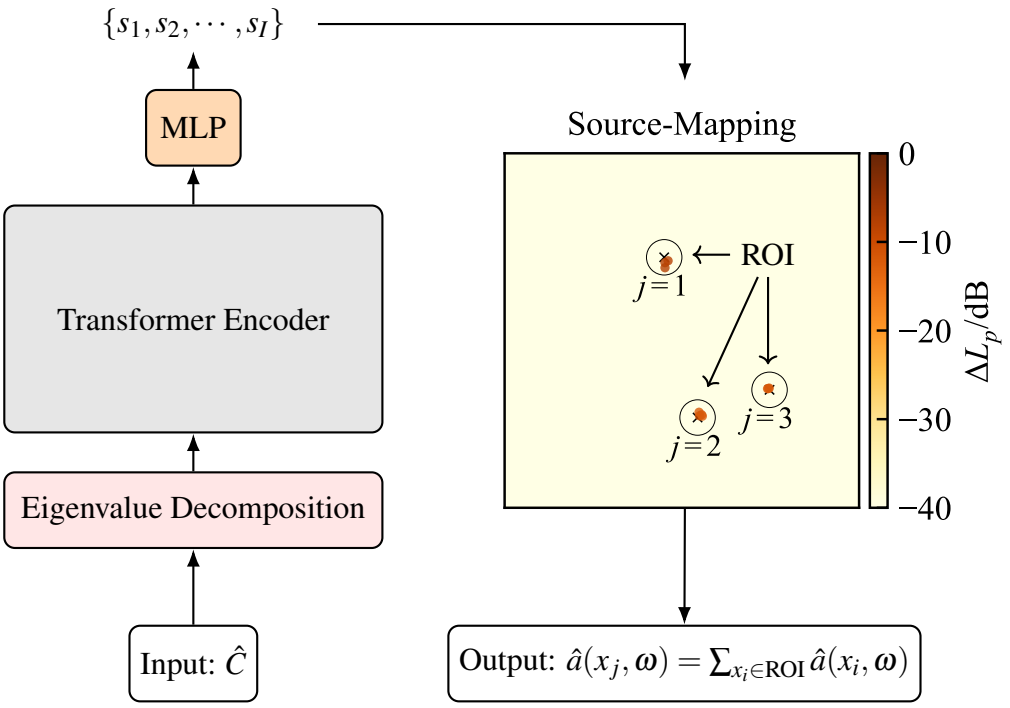

Figure 1: Model architecture of the Transformer and the source characterization procedure flow chart.

2.1. Neural Network Architecture

The Transformer architecture from [15] is employed as the deep neural network (DNN) model for solving the grid-free source characterization problem, which is depicted in Figure 1. The DNN $\mathscr{F} : \mathbb{C}^{M \times M} \rightarrow \mathscr{E}$ estimates a fixed set of $I = 10$ source characteristics based on a normalized estimate of the CSM $\hat{C} \in \mathbb{C}^{M \times M}$. The CSM is a complex-valued matrix that contains the auto- and cross-power between the microphone signals in the frequency domain and is normalized by the measured auto-power at the reference position. A crucial preprocessing step is the eigenvalue decomposition of the CSM, which leads to a set of eigenvalues and eigenvectors. Multiplication of the eigenvectors with the corresponding eigenvalues results in the eigenmodes, building the input to the Transformer encoder. In this work, the number of eigenmodes consumed by the Transformer is set to 10. After processing the eigenmodes by the Transformer encoder, a multi-layer perceptron (MLP) yields the set of estimated source characteristics.

The model parameters are optimized through supervised learning with the objective function

$$L = \sum_{j=1}^{J} \sum_{\hat{\rho}_k \in \mathscr{S}_j} \left\| \rho_j - \hat{\rho}_k \right\|_2 + \lambda \left(\alpha_j - \sum_{\hat{\alpha}_k \in \mathscr{S}_j} \hat{\alpha}_k \right)^2, \tag{3}$$

where $\|\cdot\|_2$ is the euclidean norm, $\rho_{i,j} = \left\| \hat{\rho}_i - \rho_j \right\|_2$ corresponds to the spatial distance between the i-th estimation and the j-th source normalized by the aperture size of the microphone array, and $\alpha_j = a(x_j, \omega) / \sum_{j=1}^{J} a(x_j, \omega)$ represents the normalized source strength. λ weighs

the ratio of localization and source strength reconstruction errors, and in this work, $\lambda = 1$. The assignment of model estimates $\mathscr{S}_j$ to a ground-truth source is performed, as described in [15], by solving an assignment problem formulated as a linear program.

2.2. Data

Datasets

Three datasets are used in this work, which are referred to as *Synthetic*, *MIRACLE (A2)*, and *MIRACLE (R2)*. All datasets are based on measurements with a planar 64-channel microphone array under spatially and temporally stationary conditions. The microphone array data and the respective labels were created by Monte-Carlo simulation with the AcouPipe library [14]. Details on the parameters are given in Table 1.

The Synthetic dataset is used to train, validate, and test the model's performance using independent splits. The virtual anechoic environment contains monopole sources distributed around the center of the observation area following a bivariate Normal distribution. The source num-

Table 1: Datasets and their parameterization, including random variables of the Monte-Carlo simulation.

		Synthetic	MIRACLE (A2)	MIRACLE (R2)
Environmental Parameters	microphone array	Vogel's spiral ($M = 64$)		
	aperture size (m)	$d_\mathrm{a} = 1.0$	$d_\mathrm{a} = 1.47$	$d_\mathrm{a} = 1.47$
	environment	anechoic	anechoic	specular reflection
	speed of sound (m/s)	343.0	345.3	345.4
	source type	monopole	loudspeaker	loudspeaker
	observation area	$d_\mathrm{x} = d_\mathrm{y} = 0.5 d_\mathrm{a}$		
	source distance	$d_\mathrm{z} = d_\mathrm{a}$		
Processing Parameters	sample-rate	13 720 Hz	32 kHz	32 kHz
	block size	128	256	256
	overlap	50%		
Training Parameters	training size	∞	–	–
	validation size	500	–	–
	test size	10 000		
Random Variables	source numbers	$J \sim \mathscr{U}(1, 10)$		
	source positions (m)	$x_j \sim \mathscr{N}(\sigma = 0.1688 d_\mathrm{a})$		
	source strength (Pa^2)	$p^2_{\mathrm{RMS},j} \sim \mathscr{R}(\sigma_R = 5)$		
	signal length (s)	$T \sim \mathscr{U}(1, 10)$		
	signal-to-noise ratio	$\mathrm{SNR} \sim \mathscr{U}(10^1, 10^6)$		
	mic pos noise (mm)	$x_m \sim \mathscr{N}(\sigma = 1.0)$	–	–

bers are uniformly distributed between 1 and 10, and the sources' strength follows a Rayleigh distribution. Uncorrelated white noise is added to the microphone signals to enhance realism, using a uniformly distributed signal-to-noise ratio (SNR). In addition, the considered signal length T is uniformly distributed between 1 and 10 seconds, and the actual position of each microphone is disturbed following a bivariate Normal distribution.

The *MIRACLE* database is a recently published dataset containing measured room impulse responses (RIRs) from a loudspeaker at various locations under anechoic conditions [12]. Two scenarios of the MIRACLE database are used for the Monte-Carlo simulation. As explained in the subsequent section, experimental source cases with multiple sources are created by the superposition of the source signals processed with the individually collected transfer functions. The same random variables except the microphone position noise are used for the MIRACLE and synthetic datasets. Scenario R2 is based on recordings in the anechoic chamber of the TU Berlin and contains specular reflections from a ground plate. The measurement setup is depicted in Figure 2. In contrast, scenario A2 was conducted using the same environment but without the ground plate. While the MIRACLE (A2) dataset is suitable for testing the model's generalization performance concerning the domain shift between the virtual and the experimental domain, the MIRACLE (R2) dataset is used to investigate the model's generalization behavior concerning environmental mismatch.

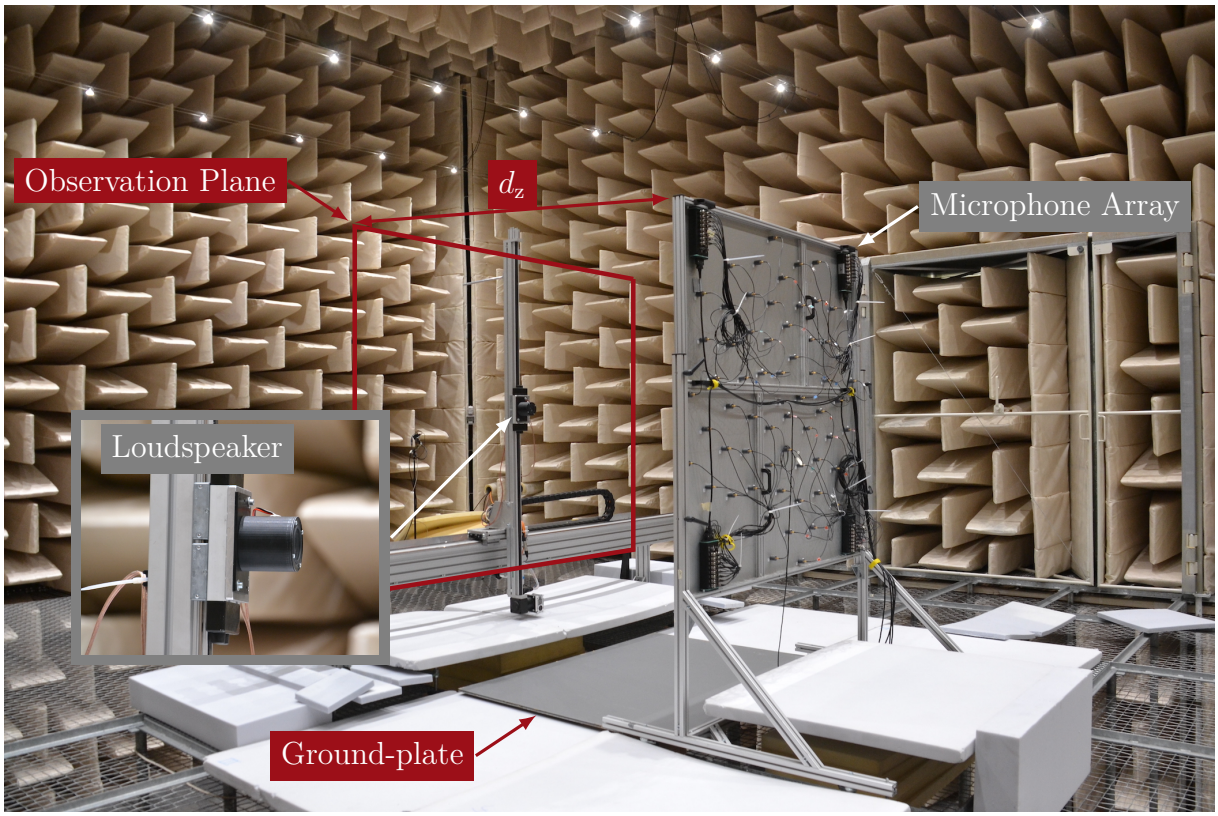

Figure 2: Experimental measurement setup R2 from the MIRACLE RIR database.

Fast generation of the CSM

Generating large amounts of synthetic data for machine learning becomes challenging if stochastic signals are involved. In real applications, the ground-truth CSM is unknown, and long measurement (and simulation) times are required to obtain a meaningful stochastic estimate $\hat{C}$. Data generation must be done before the actual training process, leading to high data storage requirements. An alternative proposed in this work is the use of a fast-to-compute approximation of the CSM, which allows the generation of the training data on the fly during training.

For a linear propagation model, the ground truth CSM can be defined as

$$C = HQH^H + N, \tag{4}$$

whereby $N \in \mathbb{C}^{M\times M}$ is the noise covariance matrix and $Q \in \mathbb{C}^{J\times J}$ is the source covariance matrix containing the source strengths $Q_{j,j}(\omega) = a(x_j,\omega)$ on the diagonal. If uncorrelated sources are assumed, $Q_{j,j'}(\omega) = 0,\ \forall j \neq j'$. $H \in \mathbb{C}^{M\times J}$ is the transfer matrix holding the transfer functions between the m-th microphone and the j-th source with respect to the reference position x_0. In the virtual measurement case, the transfer function for a monopole source is used, and

$$H_{m,j}(x_0,\omega) = \frac{H_{m,j}(\omega)}{H_{0,j}(\omega)} \in \mathbb{C}, \tag{5}$$

$$= \frac{r_{0,j}}{r_{m,j}} e^{-\mathrm{i}k(r_{m,j}-r_{0,j})}, \tag{6}$$

with $i = \sqrt{-1}$, $r_{0,j} = \|x_0 - x_j\|_2$ and $r_{m,j} = \|x_m - x_j\|_2$. The wave number k is given by $k = \omega/c_0$, where c_0 is the speed of sound. In the experimental case (MIRACLE dataset), $H_{m,j}(\omega)$ is obtained by discrete Fourier transform (DFT) of the measured impulse responses $h_{m,j}(t)$ and $h_{0,j}(t)$ so that

$$H_{m,j}(x_0,\omega) = \frac{DFT\{h_{m,j}\}(\omega)}{DFT\{h_{0,j}\}(\omega)}, \tag{7}$$

Eq. (4) offers the advantage that the CSM is fast to compute, as the sampled source strengths and the noise strength can be used directly to simulate the CSM. However, a model trained with ground-truth CSM data is assumed not to guarantee good generalization behavior when applied to estimated CSM data from a limited number of snapshots. An alternative proposed in this work is the use of an approximation for the snapshot-deficient CSM by sampling the source covariance matrix $\hat{Q}$ and the noise covariance matrix $\hat{N}$ from a complex Wishart distribution $\mathscr{W}_{\mathbb{C}}$. Since the Monte-Carlo simulation considers sound sources emitting stationary white noise, the corresponding signal vector in the frequency domain follows a multivariate complex Normal distribution

$$X \sim \mathscr{N}_{\mathbb{C}}(\mu,\Sigma) \tag{8}$$

with covariance-matrix $\Sigma \in \mathbb{C}^{J\times J}$ and mean value $\mu = 0$. The signal matrix $X = \{p(x_j,x_0,\omega)^{(k)}\} \in \mathbb{C}^{J\times K}$ contains the complex-valued sound pressure at the reference mi-

crophone for the j-th source and the k-th snapshot and

$$\hat{Q} = \frac{1}{K} X X^H. \tag{9}$$

According to [24], the Wishart distribution is the joint distribution of the sample covariance matrix of a set of random variables following a multivariate Normal distribution, and $\hat{Q}$ follows the former. This fact can be exploited to directly sample the elements of the source covariance matrix from the Wishart distribution so that

$$K\hat{Q} \sim \mathcal{W}_{\mathrm{C}}(K, Q), \tag{10}$$

which only requires to sample $\frac{1}{2}J(J+1)$ values instead of JK. Similarly, the noise covariance matrix $\hat{N}$ can be sampled from a Wishart distribution, whereby $\Sigma \in \mathbb{C}^{M\times M}$ is the identity matrix scaled by the noise variance in case of uncorrelated sensor noise. Finally, $\hat{N}$ and $\hat{Q}$ are used to compute $\hat{C}$ by means of Eq. 4.

2.3. Training and Evaluation

Three models were trained on data from a single frequency, while two models were optimized via frequency-invariant training. The latter considers the CSMs from multiple frequencies within a specific frequency range, and the model was optimized and evaluated on each frequency individually. Data from multiple source cases was shuffled before building batches for stochastic optimization. The models are summarized in Table 2. Due to the different aperture sizes of the microphone array in the Synthetic and the MIRACLE dataset, the Helmholtz number

$$\mathrm{He} = \frac{f \cdot d_{\mathrm{a}}}{c_0} \tag{11}$$

is used as a dimensionless representation of the frequency $f/\,\mathrm{Hz}$. A training epoch consisted of 2000 optimization steps with a batch size of 50 samples. The Adam optimization algorithm [20] was applied with a learning rate of $\beta = 0.25 \cdot 10^{-3}$ and weight decay regularization ($\eta = 10^{-5}$).

The characterization performance of each model was evaluated using the same metrics as in [15], which were originally introduced in [9]. The *specific level error* $\Delta L_{p,\mathrm{e,s}}/\,\mathrm{dB}$ is defined as the difference between the estimated and the ground-truth source strength in decibels, whereby the estimated source strength is integrated over a circular ROI centered on the ground-truth

Table 2: Trained Transformer models.

		Helmholtz Number		
Type	Epochs	Synth. (Training & Val.)	Synth. (Test)	A2 & R2 (Test)
single freq.	250	2.2	2.2	2.1
single freq.	250	4.1	4.1	4.2
single freq.	250	8.1	8.1	8.5
freq. invariant	1000	$[2,8]$	2.2, 4.1, 8.1	2.1, 4.2, 8.5
freq. invariant	1000	$[1,16]$	2.2, 4.1, 8.1	2.1, 4.2, 8.5

source position. The ROI radius was set to $0.05d_\mathrm{a}$, whereby the actual radius was shrunk if the distance between two sources was smaller than the ROI radius. The specific level error was used to quantify the correctness of the source strength reconstruction. The *inverse level error* $\Delta L_{p,\mathrm{e,i}}/\,\mathrm{dB}$ is defined as the difference between the sound pressure level (SPL) from all ROIs and the mapped SPL. Negative values indicate misplaced source strengths. Therefore, this metric quantifies the spatial accuracy. Similar to [15], the model-based CLEAN based on spatial source coherence (CLEAN-SC) method was evaluated on the test datasets for comparison.

3. RESULTS

3.1. Frequency-invariant training

Table 3 shows the test loss of the Transformer for the Synthetic and the MIRACLE datasets at three different Helmholtz numbers at the optimization step with the lowest validation loss. For synthetic microphone array data, the test loss is similar between all models, whereby the frequency-invariant models perform slightly worse than the single-frequency models. This performance loss is most considerable at $\mathrm{He} = 2.0$. Nevertheless, the performance loss is small and comes with the advantage of a universally applicable model. Although the frequency-invariant models were trained four times longer than the single-frequency models, the effective number of source cases per frequency is still lower for the frequency-invariant models. Low frequencies are particularly underrepresented in the training data, which could explain the performance loss. This issue can be addressed by balancing the frequency occurrence or weighting the loss function according to the frequency [17]. In contrast to the promising results for synthetic data, the model performance on the experimental data from the MIRACLE dataset is significantly worse, particularly at low frequencies and when the model was trained on multiple frequencies. The latter is likely due to the more significant number of training epochs, which causes the frequency-invariant model to overfit on the synthetic data particularly.

The results indicate that frequency-invariant training may be a viable approach if successful strategies can be found to avoid overfitting the model on synthetic data and if the training data is balanced in frequency representation. Furthermore, feeding the frequency as an additional input to the model, as done in [7], could improve performance.

Table 3: Test loss for the Transformers on the Synthetic and MIRACLE datasets A2 & R2. The test loss was evaluated for the training iteration with the lowest validation loss. The lowest test loss for each Helmholtz number and dataset is marked in bold.

	Synthetic			MIRACLE (A2)			MIRACLE (R2)		
He	$L_\mathrm{test}^\mathrm{single}$	$L_\mathrm{test}^\mathrm{He=[2,8]}$	$L_\mathrm{test}^\mathrm{He=[1,16]}$	$L_\mathrm{test}^\mathrm{single}$	$L_\mathrm{test}^\mathrm{He=[2,8]}$	$L_\mathrm{test}^\mathrm{He=[1,16]}$	$L_\mathrm{test}^\mathrm{single}$	$L_\mathrm{test}^\mathrm{He=[2,8]}$	$L_\mathrm{test}^\mathrm{He=[1,16]}$
8	**0.19**	0.20	0.21	**0.24**	0.27	**0.24**	**0.29**	0.32	**0.29**
4	**0.20**	**0.20**	0.24	**0.39**	0.49	0.47	**0.53**	0.58	0.56
2	**0.27**	0.30	0.33	**0.91**	1.32	1.17	**1.13**	1.94	1.81

3.2. Generalization

This section considers only the frequency-invariant model trained on data with Helmholtz numbers in the range between He = 1 and He = 16. The source case with the largest difference between the test loss on the Synthetic and MIRACLE (R2) dataset was selected to obtain a qualitative understanding of the model's generalization capabilities. Figure 3 and Figure 4 depict the source map for Helmholtz number four and two for the model-based CLEAN-SC and the Transformer method. The evaluated metrics are given in Table 4. CLEAN-SC suffers from the low spatial resolution but yields comparable results across the virtual and the experimental domain for He = 4, although the reflections from the ground plate cause CLEAN-SC to miss the two weakest sources on the MIRACLE (R2) source case ad to significantly overestimate the strongest source. On the other hand, the Transformer struggles to reconstruct the sources correctly for the MIRACLE (A2) dataset but almost perfectly maps the sources for the Synthetic source case.

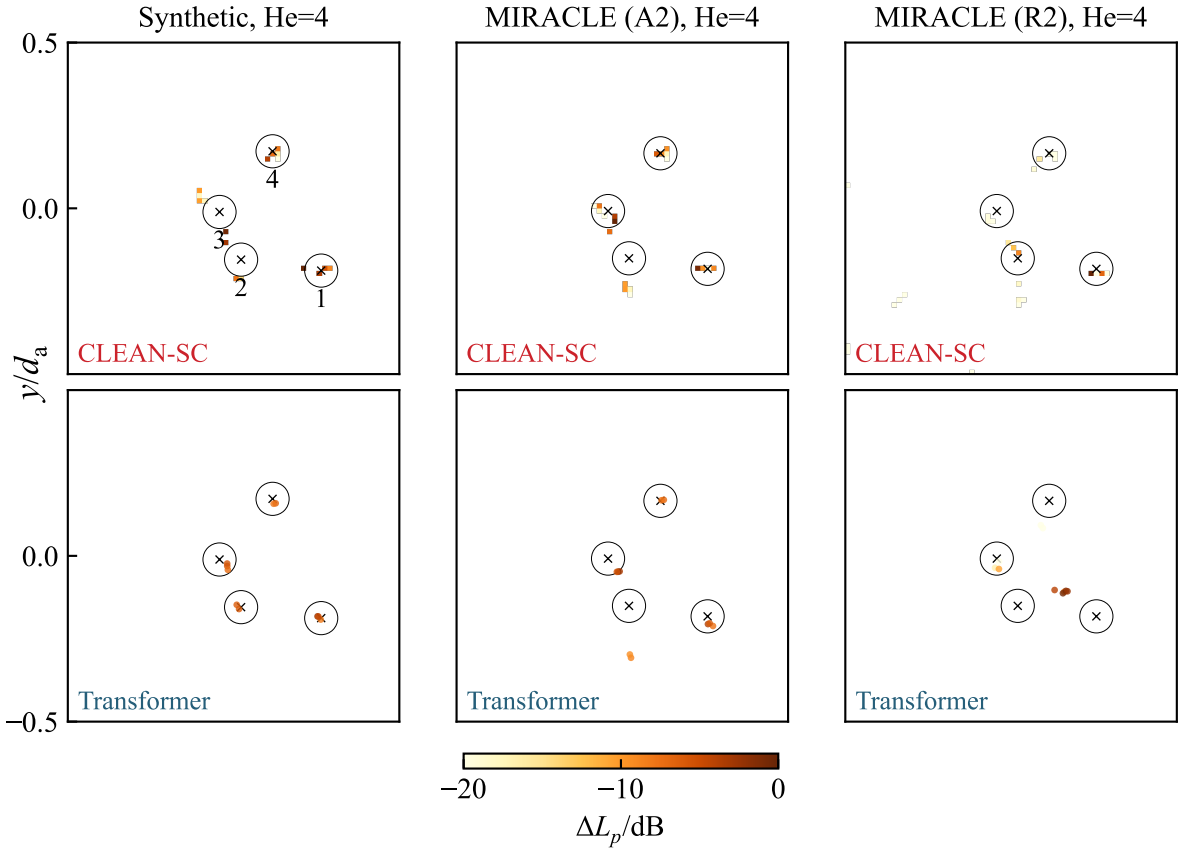

Figure 3: Source mapping for He = 4 *for the model-based CLEAN-SC (upper row) and the Transformer (lower row). The actual source positions and ROIs are marked by black crosses and circles. This source case was selected due to the largest difference in test loss between the Synthetic and MIRACLE (R2) test datasets.* ΔL_p *is the SPL relative to the SPL of the strongest ground-truth source.*

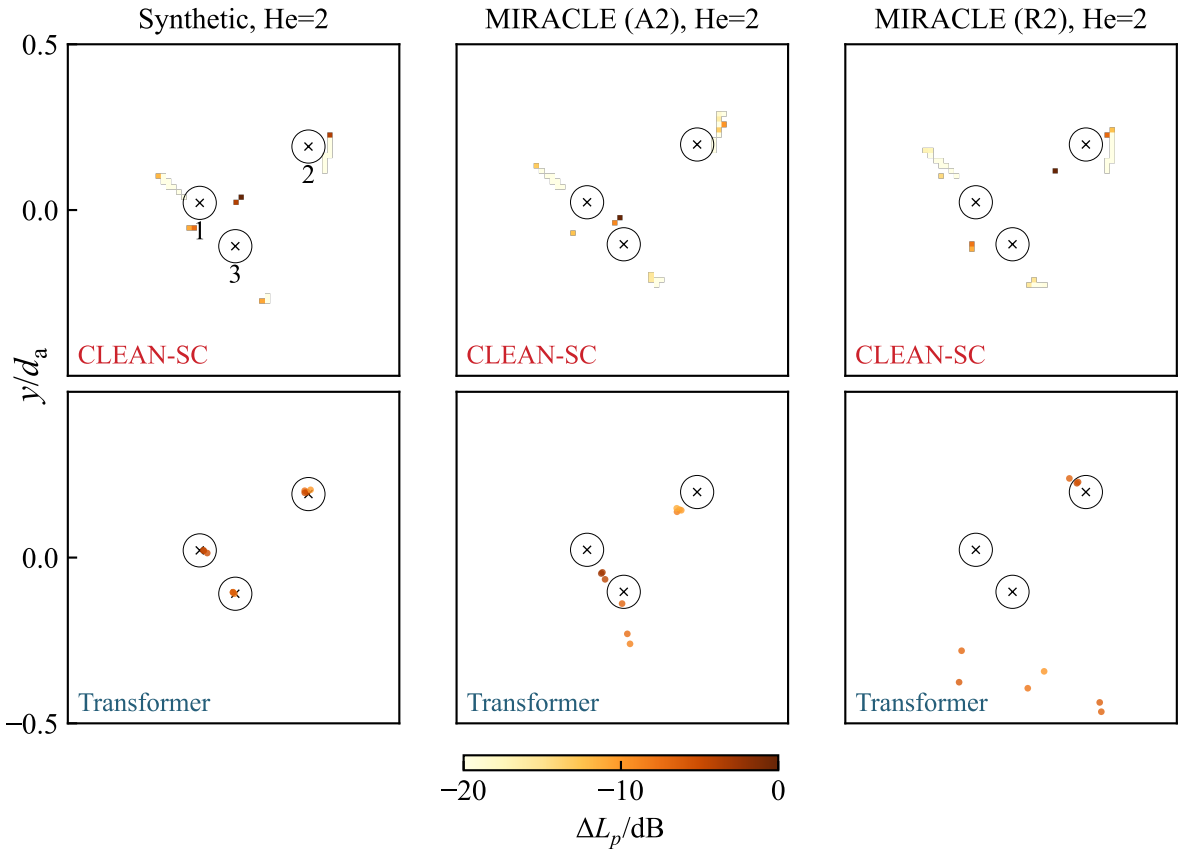

Figure 4: Source mapping for He = 2 *for the model-based CLEAN-SC (upper row) and the Transformer (lower row). The actual source positions and ROIs are marked by black crosses and circles. This source case was selected due to the largest difference in test loss between the Synthetic and MIRACLE (R2) test datasets.* ΔL_p *is the SPL relative to the SPL of the strongest ground-truth source.*

At He = 2, CLEAN-SC cannot correctly map the actual sources, which is due to the low spatial resolution obtained with CB, while the Transformer performs significantly better on the Synthetic dataset. As stated in the introduction, similar findings have been reported in the literature. However, this performance advantage does not translate to the MIRACLE datasets, where the Transformer fails to map the sources correctly. This discrepancy in performance between the Synthetic and MIRACLE datasets shows that the model was overfitted on the Synthetic dataset, and feature representations that are not generalizable to the experimental data were learned. The source mapping result for He = 8 can be found in the appendix A.

The previous observations are statistically verified in the following. For this purpose, the metrics described in Sec. 2.3 were evaluated using the respective test datasets. Figure 5 shows a histogram of the specific level error over all sound sources for the three Helmholtz numbers. An additional histogram shows the percentage of ROIs with an integrated SPL of 0 dB, meaning that no source was found inside the ROI. It is worth noting that CLEAN-SC leads to considerably less sparse source mappings than the Transformer and, therefore, has a higher chance of

detecting sources. Figure 6 shows the histogram of the inverse level error, depending on the Helmholtz number and method. An additional histogram shows the percentage of source cases where sound energy is mapped outside the ROIs.

Figure 5 reveals a performance degradation on experimental data even for the CLEAN-SC method. This is expressed by the fact that a reconstruction error of $\pm 0.5\,\mathrm{dB}$ occurs significantly less frequently for experimental data. Instead, a broader distribution of the specific level error indicates that the accuracy of the level reconstruction decreases for experimental data. Nevertheless, given the percentage of undetected sources, which is approximately constant over the different datasets, it is evident that CLEAN-SC is robust to different environmental conditions. Given the inverse level error in Figure 6, it is seen that a change from the virtual to the experimental domain does not affect the localization accuracy of CLEAN-SC significantly.

For the Transformer, the performance over the individual dataset largely depends on the frequency. For He=8, the specific level error distribution is only marginally affected by the domain shift, indicating that the model generalizes well to the experimental data at this particular frequency. However, for He=4 and He=2, the Transformer outperforms CLEAN-SC on the Synthetic dataset but fails to generalize to the experimental data. Interestingly, the performance on the experimental dataset with similar environmental conditions to the virtual domain (A2) is comparable to that of CLEAN-SC. However, the performance on the more challenging dataset (R2) is significantly worse, leading to several undetected and misplaced sources. In contrast

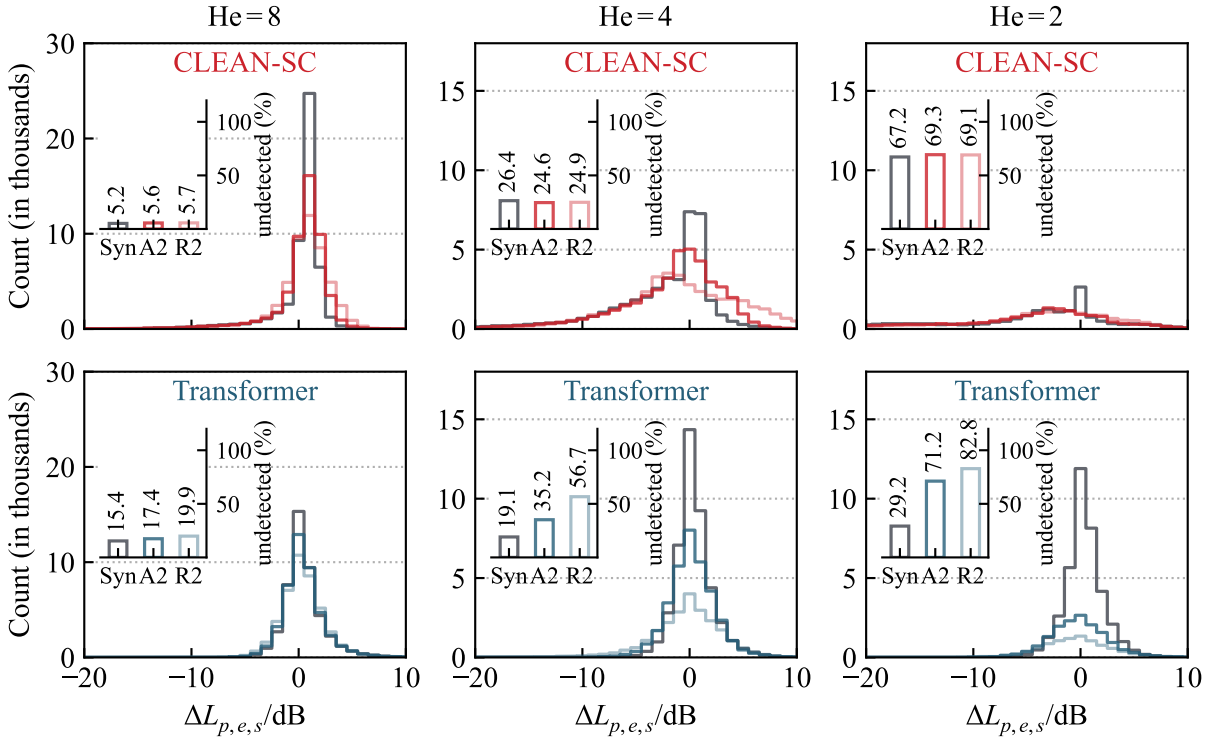

Figure 5: Histogram of the specific level error for the Transformer (lower) and CLEAN-SC (upper) for the Synthetic and MIRACLE datasets. An additional histogram shows the percentage of ROIs with resulting in a reconstructed SPL of 0 dB.

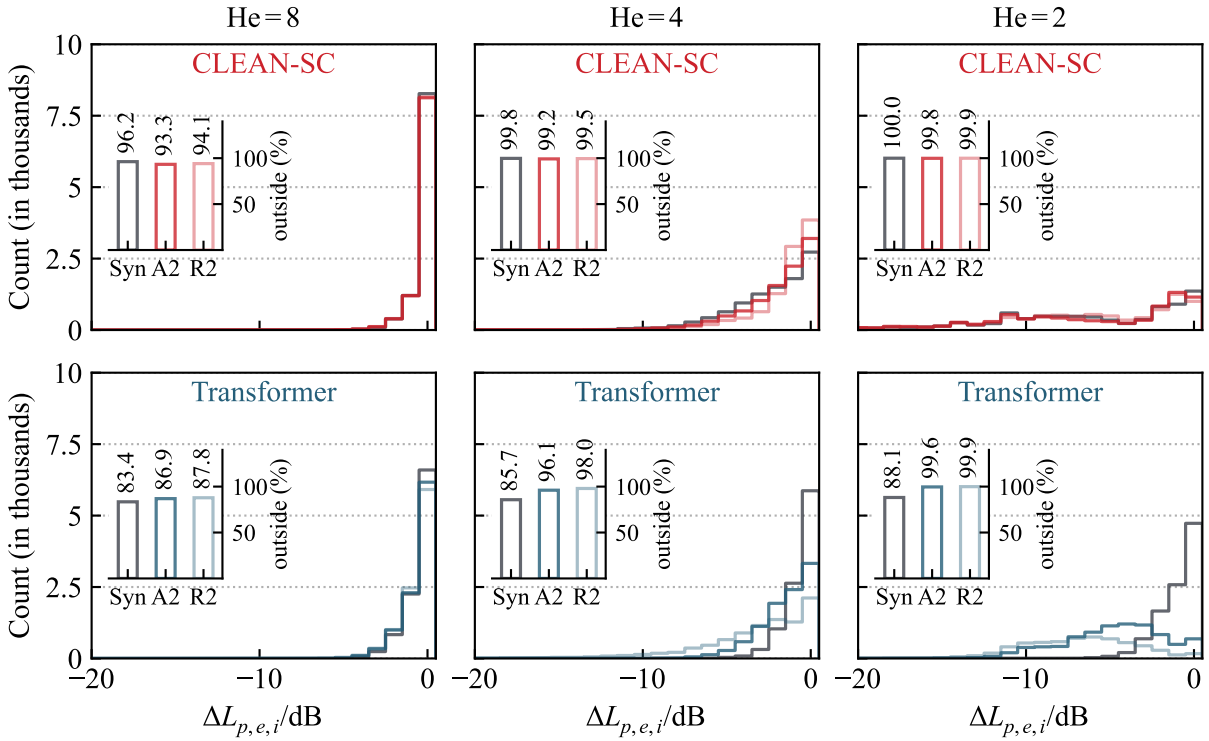

Figure 6: Histogram of the inverse level error for the Transformer (lower) and CLEAN-SC (upper) for the Synthetic and MIRACLE datasets. An additional histogram shows the percentage of source cases where source contributions are mapped outside the ROIs.

to CLEAN-SC, the inverse level error for the Transformer changes with the dataset, indicating that the localization accuracy is strongly affected by the environmental conditions.

The results indicate that the Transformer cannot yet generalize to experimental data using a naive virtual training strategy, particularly at low frequencies. Similar results can likely be observed for other data-driven methods, which is worth investigating in the future. Still, the Transformer's performance on the Synthetic dataset is promising, indicating that the model can learn to map sources correctly.

4. CONCLUSION

It was demonstrated that a single data-driven model architecture can be blindly trained with microphone array data at multiple frequencies and that the frequency-invariant model achieves comparable source characterization performance on synthetic data compared to models trained with data from only a single frequency. Although the required number of training epochs increases with the frequency range under consideration, the potential savings in training time compared to single-frequency models are considerable. Four times the number of epochs were needed to train a frequency-invariant model over a frequency range of four octave-bands. The experiments regarding the generalization of the frequency-invariant Transformer model confirmed the outstanding source characterization performance on synthetic data. However, the

performance for experimental data not part of the training decreases, especially for low frequencies. The performance loss is particularly drastic when the test data includes environmental influences such as a strong specular reflection. While this finding is not surprising, it indicates the need for research on improving experimental generalization. This could be achieved, for example, by utilizing sophisticated acoustic simulation techniques [26] and by exploring recent learning techniques from domain adaptation and generalization [30].

REFERENCES

[1] E. J. Arcondoulis, Q. Li, S. Wei, Y. Liu, and P. Xu. "Experimental validation and performance analysis of deep learning acoustic source imaging methods." *28th AIAA/CEAS Aeroacoustics Conference, 2022*, (June), 2022. doi:10.2514/6.2022-2852.

[2] 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.

[3] F. Chen, Y. Xiao, L. Yu, L. Chen, and C. Zhang. "Extending FISTA to FISTA-Net: Adaptive reflection parameters fitting for the deconvolution-based sound source localization in the reverberation environment." *Mechanical Systems and Signal Processing*, 210, 111130, 2024. doi:10.1016/j.ymssp.2024.111130.

[4] L. Feng, M. Zan, L. Huang, and Z. Xu. "A double-step grid-free method for sound source identification using deep learning." *Applied Acoustics*, 201, 109099, 2022. doi:10.1016/j.apacoust.2022.109099.

[5] A. Goudarzi, C. Spehr, and S. Herbold. "Expert decision support system for aeroacoustic classification from deconvolved beamforming maps." *AIAA AVIATION 2020 FORUM*, 2020. doi:10.2514/6.2020-2610.

[6] 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.

[7] E. Grinstein, V. W. Neo, and P. A. Naylor. "Dual input neural networks for positional sound source localization." *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1), 32, 2023. doi:10.1186/s13636-023-00301-x.

[8] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin. "A Survey of Sound Source Localization with Deep Learning Methods." *The Journal of the Acoustical Society of America*, 152(1), 107–151, 2022. doi:10.1121/10.0011809.

[9] 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.

[10] G. Jekateryńczuk and Z. Piotrowski. "A Survey of Sound Source Localization and Detection Methods and Their Applications." *Sensors*, 24(1), 68, 2023. doi:10.3390/s24010068.

[11] C. Kayser, A. Kujawski, and E. Sarradj. "A fast data-driven method for inverse microphone array signal processing." *JASA Express Letters*, 3(4), 042401, 2023. doi:10.1121/10.0017882.

[12] A. Kujawsk, A. J. R. Pelling, and E. Sarradj. "MIRACLE – A Microphone Array Impulse Response Dataset for Acoustic Learning." *EURASIP Journal on Audio, Speech, and Music Processing*. doi:10.1186/s13636-024-00352-8.

[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] A. Kujawski, A. J. R. Pelling, S. Jekosch, and E. Sarradj. "A framework for generating large-scale microphone array data for machine learning." *Multimedia Tools and Applications*, 83(11), 31211–31231, 2023. doi:10.1007/s11042-023-16947-w.

[15] A. Kujawski and E. Sarradj. "Fast grid-free strength mapping of multiple sound sources from microphone array data using a Transformer architecture." *The Journal of the Acoustical Society of America*, 152(5), 2543–2556, 2022. doi:10.1121/10.0015005.

[16] 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.

[17] S. Y. Lee and S. Lee. "Acoustic Source Localization for a Single Point Source using Convolutional Neural Network and Weighted Frequency Loss." In *Proceedings of the Inter-Noise Conference*. 2020.

[18] Q. Li, E. J. Arcondoulis, S. Wei, P. Xu, and Y. Liu. "Robustness analysis and experimental validation of a deep neural network for acoustic source imaging." *Mechanical Systems and Signal Processing*, 216, 111477, 2024. doi:10.1016/j.ymssp.2024.111477.

[19] T. Lobato, R. Sottek, and M. Vorländer. "Deconvolution with neural grid compression: A method to accurately and quickly process beamforming results." *The Journal of the Acoustical Society of America*, 153(4), 2073–2089, 2023. doi:10.1121/10.0017792.

[20] I. Loshchilov and F. Hutter. "Decoupled Weight Decay Regularization." In *Proceedings ofthe ICLR*. New Orleans, USA, 2019.

[21] 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.

[22] 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*, volume 10. Springer Vienna, 2019. ISBN 0-12-345678-9. doi:10.1007/s13272-019-00383-4.

[23] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. "A unifying view on dataset shift in classification." *Pattern Recognition*, 45(1), 521–530, 2012. doi:10.1016/j.patcog.2011.06.019.

[24] D. K. Nagar and A. K. Gupta. "Expectations of Functions of Complex Wishart Matrix." *Acta Appl Math*, 113, 265–288, 2011. doi:10.1007/s10440-010-9599-x.

[25] W. G. Pinto, M. Bauerheim, and H. Parisot-Dupuis. "Deconvoluting acoustic beamforming maps with a deep neural network." In *Proceedings of the Inter-Noise Conference*, pages 5397–5408. Institute of Noise Control Engineering, Washington, D.C., 2021. ISBN 978-1-73259-865-2. doi:10.3397/IN-2021-3084.

[26] P. Srivastava. *Realism in virtually supervised learning for acoustic room characterization and sound source localization*. Doctoral thesis, University of Lorraine, France, 2023.

[27] P. Srivastava, A. Deleforge, A. Politis, and E. Vincent. "How to (Virtually) Train Your Speaker Localizer." In *INTERSPEECH 2023*, pages 1204–1208. ISCA, 2023. doi:10.21437/Interspeech.2023-1065.

[28] P. Xu, E. J. Arcondoulis, and Y. Liu. "Acoustic source imaging using densely connected convolutional networks." *Mechanical Systems and Signal Processing*, 151, 107370, 2021. doi:10.1016/j.ymssp.2020.107370.

[29] G. Zhang, L. Geng, F. Xie, and C.-D. He. "A dynamic convolution-transformer neural network for multiple sound source localization based on functional beamforming." *Mechanical Systems and Signal Processing*, 211, 111272, 2024. doi:10.1016/j.ymssp.2024.111272.

[30] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. "Domain Generalization: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2022. doi:10.1109/TPAMI.2022.3195549.

A. Source maps

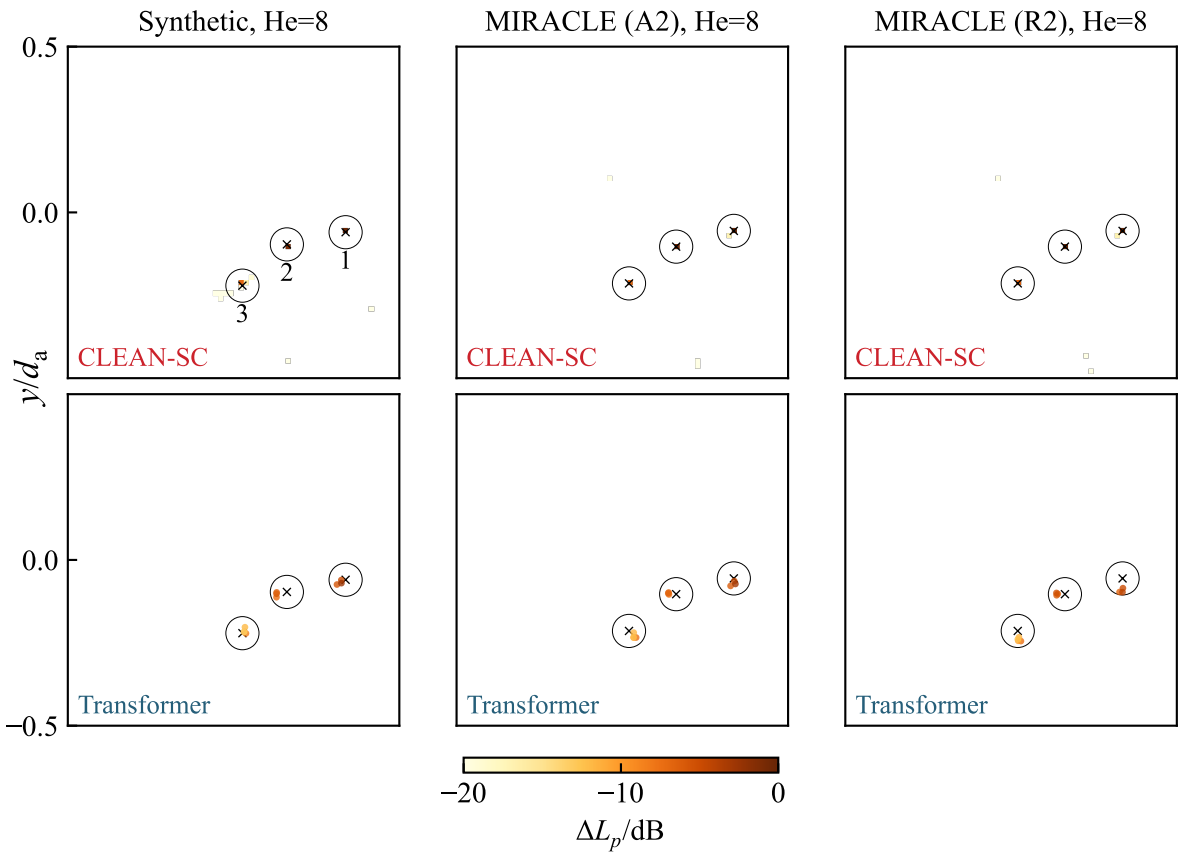

Figure 7: Source mapping for $\mathrm{He} = 8$ *for the model-based CLEAN-SC (upper row) and the Transformer (lower row). The actual source positions and ROIs are marked by black crosses and circles. This source case was selected due to the largest difference in test loss between the Synthetic and MIRACLE (R2) test datasets.* ΔL_p *is given relative to the SPL of the strongest ground-truth source.*

B. Metrics

Table 4: Reconstruction errors for the source cases depicted in Figure 7, Figure 3 and Figure 4.

	Method	Metric	Synthetic	MIRACLE (A2)	MIRACLE (R2)
Figure 7 He=8	CLEAN-SC	$L^1_{p,\mathrm{e,s}}$/dB	0.6	-0.0	-1.1
		$L^2_{p,\mathrm{e,s}}$/dB	0.7	0.0	0.1
		$L^3_{p,\mathrm{e,s}}$/dB	1.9	0.7	0.1
		$L_{p,\mathrm{e,i}}$/dB	-0.0	-0.0	-0.0
	Transf.	$L^1_{p,\mathrm{e,s}}$/dB	0.5	0.3	-0.4
		$L^2_{p,\mathrm{e,s}}$/dB	-0.6	-0.9	0.2
		$L^3_{p,\mathrm{e,s}}$/dB	2.7	3.6	3.7
		$L_{p,\mathrm{e,i}}$/dB	-0.0	0.0	0.0
Figure 3 He=4	CLEAN-SC	$L^1_{p,\mathrm{e,s}}$/dB	-2.6	-1.1	10.1
		$L^2_{p,\mathrm{e,s}}$/dB	$-\infty$	$-\infty$	4.0
		$L^3_{p,\mathrm{e,s}}$/dB	$-\infty$	2.8	-5.1
		$L^4_{p,\mathrm{e,s}}$/dB	0.6	-0.8	-1.0
		$L_{p,\mathrm{e,i}}$/dB	-4.3	-0.5	-0.2
	Transf.	$L^1_{p,\mathrm{e,s}}$/dB	0.3	-0.8	$-\infty$
		$L^2_{p,\mathrm{e,s}}$/dB	-1.2	$-\infty$	$-\infty$
		$L^3_{p,\mathrm{e,s}}$/dB	1.0	1.6	-7.3
		$L^4_{p,\mathrm{e,s}}$/dB	-0.5	-0.7	$-\infty$
		$L_{p,\mathrm{e,i}}$/dB	0.0	-1.3	-13.8
Figure 4 He=2	CLEAN-SC	$L^1_{p,\mathrm{e,s}}$/dB	$-\infty$	$-\infty$	$-\infty$
		$L^2_{p,\mathrm{e,s}}$/dB	$-\infty$	$-\infty$	$-\infty$
		$L^3_{p,\mathrm{e,s}}$/dB	$-\infty$	$-\infty$	$-\infty$
		$L_{p,\mathrm{e,i}}$/dB	$-\infty$	$-\infty$	$-\infty$
	Transf.	$L^1_{p,\mathrm{e,s}}$/dB	-0.1	$-\infty$	$-\infty$
		$L^2_{p,\mathrm{e,s}}$/dB	0.5	$-\infty$	-0.7
		$L^3_{p,\mathrm{e,s}}$/dB	0.2	-3.3	$-\infty$
		$L_{p,\mathrm{e,i}}$/dB	0.0	-10.1	-5.7