forum acusticum 2026
8-12 September · Graz, Austria

TOWARDS MICROPHONE ARRAY GEOMETRY AGNOSTIC SOUND SOURCE CHARACTERIZATION: A MESSAGE PASSING NEURAL NETWORK APPROACH

Siavash Zaid[1], Adam Kujawski[1], Ennes Sarradj[1]

[1] *Fachgebiet Technische Akustik, Technische Universität Berlin, Germany*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Sound Source Characterization (SSC) with microphone arrays has seen significant advances through deep learning. However, most data-driven methods are trained on data from a single, fixed microphone array geometry, making the resulting models geometry-specific and difficult to transfer to other configurations. Moreover, many architectures do not explicitly use sensor position information. To overcome this limitation, a Transformer architecture that uses a Message Passing Neural Network for array-agnostic SSC is proposed. The proposed model operates on a graph constructed from spatial and spectral information of the microphones, from which the location and strength of the dominant sound source in the region of interest is estimated. The model is evaluated across multiple array geometry classes, such as spiral, grid and randomized layouts, and different numbers of channels outside the training distribution, demonstrating robust generalization where a fixed-geometry baseline fails entirely. Furthermore, training with a variable number of microphones is shown to improve performance. To the authors' knowledge, this is the first work on grid-free SSC generalizing across large-scale array geometries of up to 128 microphones.

Keywords: *sound source characterization, microphone array processing, graph neural network, deep learning, microphone array agnostic modeling*

1. Introduction

Sound Source Characterization (SSC) using microphone arrays, that is, the localization and quantification of sound sources in a region of interest, is a fundamental task in acoustic signal processing, with applications ranging from communication systems to acoustic testing. In practice, microphone arrangements vary significantly in number of channels and spatial distribution depending on the specific application. Although model-based microphone array methods are inherently array-agnostic as they rely on a simplified sound propagation model, their performance can degrade substantially when the actual propagation conditions differ from this assumption, for example in reverberant environments [1]. In recent years, deep learning methods have emerged as an alternative to model-based methods, achieving superior reconstruction accuracy without the need to define an explicit sound propagation model. However, these approaches are typically trained on a fixed array geometry and architecturally constrained to a fixed microphone count, meaning that they cannot be applied directly to arbitrary microphone arrangements. Instead, costly retraining or domain adaptation is required, limiting their practical applicability [2]. Recent work has explored several directions toward data-driven array-agnostic SSC, yet fundamental limitations in scale and generalization persist. A natural starting point is to include spatial information alongside acoustic measurements in the model input. In [3], [4], microphone positions are concatenated to acoustic features, demonstrating improved robustness to unseen array geometries. However, these approaches rely on deep learning architectures with a fixed input dimensionality and are therefore limited to a specific number of microphones. An alternative approach for small microphone arrays of up to eight channels has been used in [5], aggregating features over pairs of microphones, resulting in a fixed-size representation. However, their approach relies on model-based features and reduces localization to peak selection over a fixed set of candidate directions, limiting spatial resolution. Graph-based neural networks offer a more structured approach to processing signals from variable-sized arrays as they impose no constraint on the problem dimension and are permutation-invariant by construction [6]. In [7], the microphone array is modeled as a fully connected graph, representing microphones as nodes with their positions as node features and respective

***Corresponding author**: siavash.zaid@campus.tu-berlin.de.*

cross-correlations as edge features. A similar graph representation is adopted in [8], using pairwise spatial likelihood maps as edge features with microphone positions supplied as auxiliary input. However, both rely on intermediate representations rather than directly performing grid-free SSC. [7] regresses intermediate Time-difference of Arrival (TDoA) values from which the Direction of Arrival (DoA) is recovered analytically, while [8] refines precomputed spatial likelihood maps rather than predicting source location directly from raw spectral features. As with the preceding methods, both are evaluated on small arrays of up to five and seven microphones, respectively. In addition, none of the mentioned publications are concerned with the prediction of the source strength.

This work proposes an architecture for grid-free SSC that operates directly on auto- and cross-power spectra of the microphones and is agnostic to microphone placement and the number of channels, as it learns a permutation-invariant mapping from microphone locations and spectra to source characteristics. The proposed method is evaluated on large-scale planar arrays with up to 128 microphones and demonstrates domain generalization across different microphone geometry classes. In addition, training with a varying number of microphones is shown to improve generalization and to yield significantly better performance than training with a fixed number of microphones, even when evaluated at the same channel count.

2. Theory

2.1. Propagation Model

It is assumed that J uncorrelated and spatially stationary sources are observed by a planar array with M microphones. In the noise-free case and under a linear propagation model, the complex sound pressure at an angular frequency ω at microphone m is given by

$$p_m(\omega) = \sum_{j=1}^{J} h_{mj}(\omega) q_j(\omega) \tag{1}$$

where $q_j(\omega)$ is the complex source amplitude of source j and $h_{mj}(\omega)$ is the transfer function from source j to microphone m. Collecting the microphone pressures in a vector yields

$$\mathbf{p}(\omega) = \mathbf{H}(\omega)\mathbf{q}(\omega), \tag{2}$$

with $\mathbf{p}(\omega) \in \mathbb{C}^M$, $\mathbf{q}(\omega) \in \mathbb{C}^J$, and $\mathbf{H}(\omega) \in \mathbb{C}^{M\times J}$. If microphone m_0 is chosen as a reference position, the strength of source j can be expressed by the auto-power it induces at that sensor,

$$a_j(\omega) = |h_{m_0 j}(\omega)|^2 \, \mathbb{E}\left\{|q_j(\omega)|^2\right\}, \tag{3}$$

where $\mathbb{E}\{\cdot\}$ is the expectation value. Normalizing $a_j(\omega)$ by $\sum_{k=1}^{J} a_k(\omega)$ gives the relative source strength.

2.2. Cross-Spectral Matrix

The Cross-Spectral Matrix (CSM) of the microphone signals is defined as

$$\mathbf{C}(\omega) = \mathbb{E}\left\{\mathbf{p}(\omega)\mathbf{p}^{\mathrm{H}}(\omega)\right\}, \tag{4}$$

where $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose. The diagonal entries correspond to the microphone auto-power spectra, whereas the off-diagonal entries contain pairwise cross-power spectra. Inserting Eq. (2) leads to

$$\mathbf{C}(\omega) = \mathbf{H}(\omega)\mathbf{S}(\omega)\mathbf{H}^{\mathrm{H}}(\omega), \tag{5}$$

with $\mathbf{S}(\omega) = \mathbb{E}\left\{\mathbf{q}(\omega)\mathbf{q}^{\mathrm{H}}(\omega)\right\}$. For uncorrelated sources, $\mathbf{S}(\omega)$ is diagonal.

3. Method

Given a planar microphone array of M microphones, the proposed method

$$\mathcal{F} : \mathbb{R}^{2\times M} \times \mathbb{C}^{M\times M} \to \mathbb{R}^2 \times \mathbb{R}, \quad M \in \mathbb{N} \tag{6}$$

is supposed to predict the spatial location of the strongest source $\mathbf{r} \in \mathbb{R}^2$ in a planar focus area and its relative source strength $a \in \mathbb{R}$, based on the positions of the microphones $\mathbf{R} \in \mathbb{R}^{2\times M}$ and the CSM according to Eq. (4). The prediction task is restricted to the strongest of multiple sources as in [9]. This keeps the focus on investigating the viability of the proposed approach, with multi-source prediction being a natural extension through a different prediction head and training objective [10]. $\mathbf{R}$ and $\mathbf{C}(\omega)$ are represented as a fully connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $\mathbf{v}_i \in \mathcal{V}$ holds information from a single microphone and each edge $\mathbf{e}_{ij} \in \mathcal{E}$ encodes the spatial and acoustic relationship between $\mathbf{v}_i$ and $\mathbf{v}_j$. Nodes $\mathbf{v}_i$ and edges $\mathbf{e}_{ij}$ are described by feature vectors $\mathbf{x}_i \in \mathbb{R}^4$ and $\mathbf{e}_{ij} \in \mathbb{R}^3$, respectively. Node features $\mathbf{x}_i$ comprise the auto-power C_{ii}, the radial distance r_i from the array center, and the polar angle encoded as $(\sin\theta_i, \cos\theta_i)$ to avoid the discontinuity arising from direct angle representation, while edge features $\mathbf{e}_{ij}$ contain the pairwise distance d_{ij} and the real and imaginary parts of the cross-spectrum C_{ij}. Following [10], the CSM is normalized by the auto-power of the microphone closest to the center of the array, preserving relative power ratios and phase differences while removing absolute magnitude.

3.1. Model Architecture

The proposed model architecture handles a variable number of microphones and is permutation-invariant with respect to the ordering of the channels. It consists of three sequential modules: a Message Passing Neural Network (MPNN), a Transformer encoder, and a Multilayer Perceptron (MLP) for prediction of the source characteristics. Figure 1 illustrates the computation graph alongside a detailed view of the MPNN as its central component. The MPNN operates on graph $\mathcal{G}$ to produce one feature vector $\mathbf{h}_i$ per microphone node $\mathbf{v}_i$. Before the first of four message passing layers,

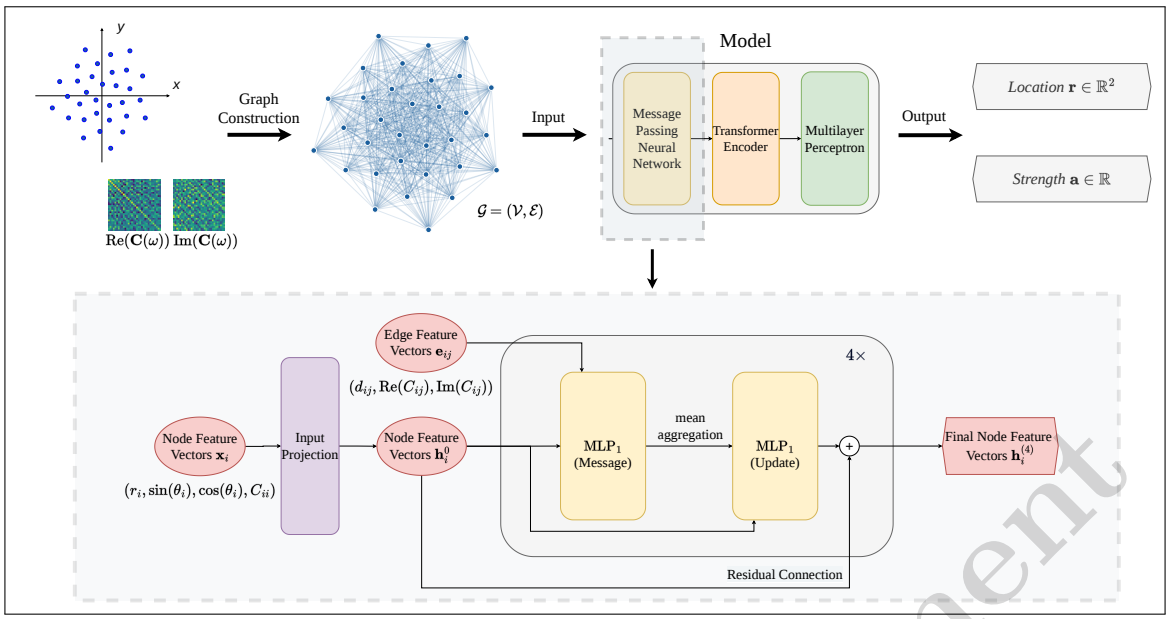

Figure 1: End-to-end pipeline of the proposed method. The microphone positions and CSM are first encoded into a fully connected graph $\mathcal{G}$, which is processed by the MPNN to produce one node feature vector per microphone. The resulting vector sequence is contextualized by a Transformer encoder and averaged into a fixed-size representation, from which an MLP predicts source location $\mathbf{r} \in \mathbb{R}^2$ and strength $a \in \mathbb{R}$. The lower panel details the internal structure of the MPNN, consisting of an input projection and 4× message passing layers with residual connections.

the feature vector $\mathbf{x}_i$ of each microphone is linearly projected to dimension $D_\mathrm{n} = 64$, yielding the initial node representation $\mathbf{h}_i^{(0)}$. Following the message passing framework of [11], each layer l iteratively updates the node feature vectors $\mathbf{h}_i^{(l)}$ by aggregating messages from all neighboring nodes $\mathcal{N}(i)$,

$$\mathbf{h}_i^{(l+1)} = \mathrm{MLP}_2\left(\mathbf{h}_i^{(l)}, \bigoplus_{j \in \mathcal{N}(i)} \mathrm{MLP}_1\left(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}, \mathbf{e}_{ij}\right)\right) \tag{7}$$

where MLP_1 computes a message for each microphone pair $(\mathbf{v}_i, \mathbf{v}_j)$ from their respective feature vectors $\mathbf{h}_i^{(l)}$ and edge features $\mathbf{e}_{ij}$, while MLP_2 produces the updated node feature vectors $\mathbf{h}_i^{(l+1)}$ by processing the concatenation of the current feature vector $\mathbf{h}_i^{(l)}$ and all aggregated messages. MLP refers to a learned nonlinear function implemented as a two-layer fully connected network with Rectified Linear Unit (ReLU) activations and Dropout. Mean aggregation $\bigoplus$ is used to prevent message magnitudes from scaling with M since the graph is fully connected. Each layer further includes a residual connection, adding the previous node representation $\mathbf{h}_i^{(l)}$ to the updated one $\mathbf{h}_i^{(l+1)}$. The resulting M node feature vectors $\mathbf{h}_i^{(4)}$ are passed to a Transformer encoder derived from the Vision Transformer (ViT) architecture [12], capturing dependencies between all node feature vectors through self-attention across the full array. The encoder consists of 6 layers with 8 attention heads. The transformed vectors are then averaged into a single fixed-size vector, which is passed to an MLP with a hidden layer of dimension 256, followed by two output branches predicting location $\mathbf{r} \in \mathbb{R}^2$ and strength $a \in \mathbb{R}$.

3.2. Baseline Model Architecture

A similar state-of-the-art architecture without MPNN is utilized as a baseline to demonstrate the effectiveness and generalization capabilities of the MPNN approach. The baseline follows the ViT-derived architecture of [10], an earlier publication of the authors, which receives a sequence of CSM eigenmodes as input features. The eigenmodes are the eigenvectors of the CSM scaled by their corresponding eigenvalues. Since the eigenmodes do not carry explicit spatial information about the microphone locations and their representation depends on the ordering of the microphone channels, the baseline is expected to be sensitive to channel permutations and changes of the microphone arrangement. With 5.8 M trainable parameters compared to 550 k for the proposed model, the baseline is considerably larger. This discrepancy can be attributed to the subsequent processing of the Transformer's output. The baseline concatenates all transformed feature vectors as input to the MLP, whereas the proposed method applies average pooling over the sequence, resulting in substantially lower dimensionality.

4. Experiments

Several classes of planar array geometries can be distinguished, with suitability depending on the application, including Vogel's Spiral [13], circular, rectangular and random configurations [1]. Accordingly, the experi-

ments should assess not only generalization to unseen arrangements from the same class but also across geometry classes not represented during training. Another important aspect concerns the influence of channel number, which may vary considerably between applications. Consequently, it is important to assess whether a model can generalize to channel counts that differ from those seen during training. Therefore, two different experiments were conducted to evaluate the characterization performance:

***Experiment 1*:** considers a fixed number of microphones with positional variations. All training arrangements are based on Vogel's spiral with a variation in the radial density [14]. To test generalization to unseen geometries, evaluation is additionally performed on the random and grid classes further described in Section 4.1.

***Experiment 2*:** extends Experiment 1 by additionally varying the microphone count. Training is conducted on Vogel spiral arrangements following $M \sim \mathcal{U}(32, 64)$. Evaluation spans three array-size ranges: (i) below the training range $M \sim \mathcal{U}(16, 31)$; (ii) in-distribution $M \sim \mathcal{U}(32, 64)$; (iii) above the training range $M \sim \mathcal{U}(65, 128)$. A comparison with the baseline is not possible in this experiment since it requires a fixed number of microphones.

4.1. Data

The microphone array data used for the experiments were generated with the openly available AcouPipe framework [15]. Each source case contains one to four spatially stationary white noise emitters, positioned within a $1\,\mathrm{m} \times 1\,\mathrm{m}$ focus plane at a distance of $0.5\,\mathrm{m}$ from the array. The sources radiate under anechoic conditions at a single frequency of $f = 2.5\,\mathrm{kHz}$. All training and validation array geometries follow Vogel spiral arrangements [13], shown to yield Pareto-optimal properties in terms of beam width and the maximum side lobe level [14]. The evaluation data were divided into three geometry classes, of which the latter two are unseen during training:

***Spiral*:** Vogel's spiral with radial density parameter $H \sim \mathcal{U}(-1, 4)$ [14]

***Random*:** microphone positions placed uniformly over a disc with a diameter of $1\,\mathrm{m}$

***Grid*:** rectangular grid with randomly sampled row count $R \in [2, \lfloor\sqrt{M}\rfloor]$ and corresponding column count C such that $R \cdot C = M$; microphones uniformly spaced over a fixed extent such that the aperture equals $1\,\mathrm{m}$

Figure 2 illustrates the three geometry classes alongside example predictions for each.

4.2. Training

All models are trained by minimizing the mean squared error of both predictions over a batch of n source cases, defined as

$$L = \frac{1}{n} \sum_{i=1}^{n} \left(\|\mathbf{r}_i - \hat{\mathbf{r}}_i\|_2^2 + (a_i - \hat{a}_i)^2 \right), \qquad (8)$$

with $\|\cdot\|_2^2$ being the squared Euclidean norm.
All models are randomly initialized and trained using the AdamW optimizer with a learning rate of $5 \cdot 10^{-4}$, a weight decay of $1 \cdot 10^{-4}$ and gradient clipping at a maximum norm of 1.0 for 200 epochs with a batch size of 256 samples. Model performance is evaluated on localization accuracy, reported as the mean Euclidean distance between predicted and true source positions in centimeters, and strength accuracy, reported as the mean absolute error in sound pressure level in decibels. All experiments share a split of 5×10^5 training and 6.25×10^4 validation and test samples each.

5. Results and Discussion

Figure 2 shows example predictions of the proposed model for each geometry class. Figures 3 and 4 summarize localization and strength estimation errors as

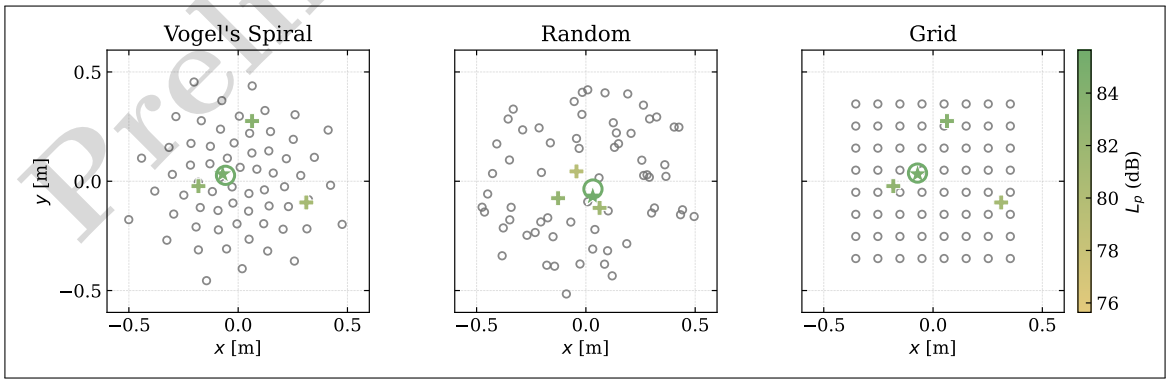

Figure 2: Example predictions of the proposed model (Experiment 2, variable-size) for each of the three evaluation geometry classes. Only the dominant source is predicted. Microphone positions are shown as grey circles. Filled markers indicate ground truth source positions colored by absolute sound pressure level L_p (dB), with stars denoting the dominant source and plus markers denoting secondary sources. Open circle markers indicate the predicted source position colored by predicted L_p.

boxplots, with boxes spanning the 25th and 75th percentiles and whiskers extending to 1.5 ×Interquartile Range (IQR).

5.1. Experiment 1: In-Distribution Performance and Geometry Generalization

Both models perform comparably on unseen spiral-type geometries, with the proposed model achieving a median localization error of 0.73 cm (IQR: 0.71 cm) against 0.58 cm (IQR: 0.73 cm) for the baseline. The large model capacity allows the baseline to learn mappings blindly for different eigenmode representations. However, the baseline does not generalize to unseen array topologies, degrading to a median error of 28 cm (IQR: 21 cm) for both unseen classes, suggesting it has learned array-specific mappings rather than transferable spatial-acoustic relationships, rendering it unusable without retraining. In contrast, the proposed model maintains median localization errors of 2.23 cm (IQR: 2.59 cm) for random arrays and 2.66 cm (IQR: 11.94 cm) for grid arrays. Strength estimation errors remain consistently low across all evaluated array topologies, with a median absolute error below 0.5 dB. The slight degradation on non-spiral geometries is expected, as spiral geometries possess significantly better spatial filtering properties than the other two geometry classes [14]. Another plausible explanation is that the spiral geometries used for training represent only a subset of possible spatial relationships between microphones. We leave the question of how training geometries should be designed for optimal domain generalization to future work.

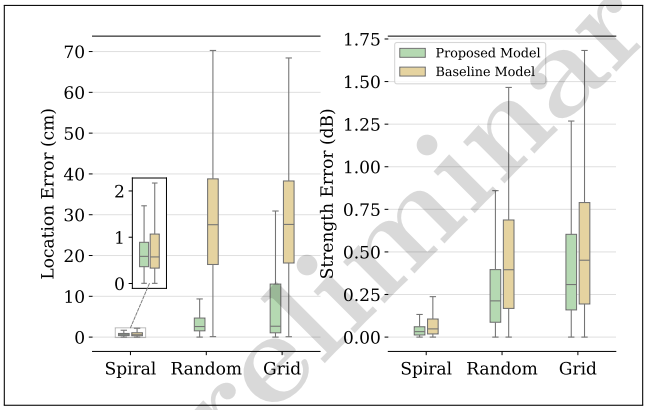

Figure 3: Localization error and source strength error of the main model and baseline across three array geometries: Vogel Spiral (in-distribution), random positions and random grid (both unseen during training), each with 64 channels. The boxes span the 25th and 75th percentiles, and the whiskers extend to 1.5 × IQR.

5.2. Experiment 2: Generalization for Different Channel Counts

For all three geometry classes, localization error decreases consistently with increasing microphone count. For spiral arrays, performance improves notably from 16+ to 32+ microphones (median: 1.22 cm, IQR: 1.69 cm to 0.82 cm, 0.91 cm) before plateauing at 64+ microphones (median: 0.8 cm, IQR: 0.85 cm). For random arrays, the largest reduction also occurs between 16+ and 32+ microphones (median: 4.67 cm, IQR: 6.66 cm to 2.51 cm, 2.89 cm), with continued improvement at 64+ (median: 1.69 cm, IQR: 1.73 cm). Grid arrays follow the same trend (median: 3.3 cm to 1.31 cm and IQR: 11.71 cm to 2.13 cm), approaching the other two geometry classes at 64+ microphones. The substantially larger IQR relative to the median at low microphone counts points to grid-specific failure modes that produce a small number of large outlier errors, consistent with grating lobes and limited usable frequency range introduced by uniform microphone spacing [16], which diminish as increasing array density provides sufficient spatial sampling to partially compensate. The results demonstrate that the proposed model generalizes well across array sizes. Both random and grid arrays approach in-distribution Spiral performance at 64+ microphones, suggesting that the degradation observed in Experiment 1 is not primarily a failure of generalization but a consequence of the suboptimal spatial sampling properties of the tested arrays. A particularly noteworthy finding emerges when comparing the fixed-size and variable-size models. For non-Spiral arrays, the variable-size model outperforms the fixed-size model from 32+ microphones onward, despite being evaluated on arrays with fewer microphones on average (32 to 64 vs. a fixed 64), suggesting that training diversity across array sizes prevents overfitting to specific array configurations and promotes robust learning. The fixed-size model retains a slight advantage on in-distribution Vogel Spirals. Strength estimation remains consistently low across all evaluated geometries and microphone counts with a median absolute error below (0.3 dB).

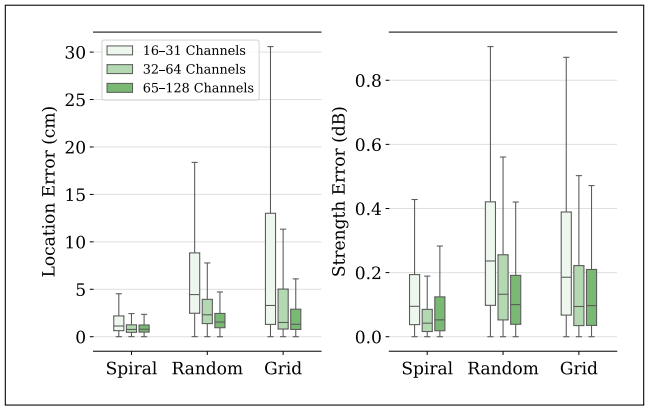

Figure 4: Localization error and source strength error of the main model across three array geometries and three microphone counts: below training range (16-31 Channels), in-distribution (32-64 Channels), and above training range (65-128 Channels). The boxes span the 25th and 75th percentiles, and the whiskers extend to 1.5 × IQR.

6. Conclusion

This work presented an architecture for SSC that generalizes across microphone array geometries and sizes without retraining, leveraging a message passing framework that enables permutation-invariant processing for arrays of arbitrary size and geometry. The performance evaluation across multiple geometry classes demonstrated competitive in-distribution localization accuracy relative to a state-of-the-art baseline, while generalizing to unseen array topologies where the baseline fails entirely. The architecture further demonstrates the ability to operate across a wide range of microphone counts outside the training distribution. Notably, training with a variable number of microphones is shown to improve out-of-distribution performance compared to fixed-size training, despite being evaluated on arrays with fewer microphones on average, suggesting that exposure to diverse array configurations promotes robust generalization. Source strength estimation is found to be largely geometry- and size-independent.

Future work should address explicit multi-source characterization and consider experimental datasets. In particular, the proposed architecture enables the use of a broad range of experimental datasets for training and domain adaptation, including datasets that were previously incompatible with fixed-geometry models.

References

[1] G. Jekateryńczuk and Z. Piotrowski, "A survey of sound source localization and detection methods and their applications," *Sensors*, vol. 24, no. 1, pp. 1–25, 2024.

[2] P.-A. Grumiaux et al., "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[3] 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*, vol. 2023, no. 1, pp. 1–12, 2023.

[4] U. Kowalk, S. Doclo, and J. Bitzer, "Geometry-aware doa estimation using a deep neural network with mixed-data input features," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1–5.

[5] J. Fan et al., "A unified geometry-aware source localization and separation framework for ad-hoc microphone array," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 725–729.

[6] J. Zhou et al., "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, no. 1, pp. 57–81, 2020.

[7] I. Jeong et al., "Micgraphnet: Microphone graph network for cross-correlation-based time delay estimation for accurate sound source localization," *Measurement*, vol. 258, no. Part E, pp. 1–11, 2026.

[8] E. Grinstein, M. Brookes, and P. A. Naylor, "Graph neural networks for sound source localization on distributed microphone networks," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[9] P. Castellini et al., "A neural network based microphone array approach to grid-less noise source localization," *Applied Acoustics*, vol. 177, no. 1, p. 107 947, 2021.

[10] 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*, vol. 152, no. 5, pp. 2543–2556, 2022.

[11] J. Gilmer et al., "Neural message passing for quantum chemistry," in *ICML'17: Proc. of the 34th International Conference on Machine Learning - Volume 70*, 2017, pp. 1263–1272.

[12] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. of the 9th International Conference on Learning Representations, ICLR 2021*, 2021, pp. 1–22.

[13] H. Vogel, "A better way to construct the sunflower head," *Mathematical Biosciences*, vol. 44, no. 3, pp. 179–189, 1979.

[14] E. Sarradj, "A generic approach to synthesize optimal array microphone arrangements," in *BeBeC, 6th Berlin Beamforming Conference, February 29 - March 1, 2016, Proc.*, 2016, pp. 1–12.

[15] A. Kujawski et al., "A framework for generating large-scale microphone array data for machine learning," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 31 211–31 231, 2024.

[16] C. Schulze, E. Sarradj, and A. Zeibig, "Characteristics of microphone arrays," in *Proc. of Internoise 2004*, 2004, pp. 1–7.