DOI: 10.61782/fa.2025.0301

DIFFUSIONRIR: ROOM IMPULSE RESPONSE INTERPOLATION USING DIFFUSION MODELS

Sagi Della Torre[1] **Mirco Pezzoli**[2] **Fabio Antonacci**[2] **Sharon Gannot**[1*]

[1] Faculty of Engineering, Bar-Ilan University, Israel
[2] Dipartimento di Elettronica, Informatica e Bioingegneria, Politecnico di Milano, Italy

ABSTRACT

Room Impulse Responses (RIRs) characterize acoustic environments and are crucial in multiple audio signal processing tasks. High-quality RIR estimates drive applications such as virtual microphones, sound source localization, augmented reality, and data augmentation. However, obtaining RIR measurements with high spatial resolution is resource-intensive, making it impractical for large spaces or when dense sampling is required. This research addresses the challenge of estimating RIRs at unmeasured locations within a room using Denoising Diffusion Probabilistic Models (DDPM). Our method leverages the analogy between RIR matrices and image inpainting, transforming RIR data into a format suitable for diffusion-based reconstruction.

Using simulated RIR data based on the image method, we demonstrate our approach's effectiveness on microphone arrays of different curvatures, from linear to semi-circular. Our method successfully reconstructs missing RIRs, even in large gaps between microphones. Under these conditions, it achieves accurate reconstruction, significantly outperforming baseline Spline Cubic Interpolation (SCI) in terms of Normalized Mean Square Error (NMSE) and Cosine Distance (CD) between actual and interpolated RIRs.

This research highlights the potential of using generative models for effective RIR interpolation, paving the way for generating additional data from limited real-world measurements.

Keywords: *Diffusion models, RIR interpolation*

1. INTRODUCTION

Room Impulse Responses (RIRs) play a critical role in audio signal processing, enabling applications such as sound source localization, virtual and augmented reality, and data augmentation for machine learning. However, measuring RIRs is resource-intensive, particularly in large or acoustically complex spaces requiring dense measurements. Simulated RIRs, while practical, often lack the accuracy and fidelity of real-world data, necessitating methods to reconstruct or interpolate RIRs at unmeasured locations.

Traditional methods for RIR reconstruction rely on mathematical models, such as compressed sensing and wave equation solutions [1–5], but these approaches often struggle with complex acoustic environments. Recent advancements leverage deep learning techniques, including Convolutional Neural Networks (CNNs) [6] and Generative Adversarial Networks (GANs), to improve reconstruction accuracy. For instance, GANs have shown promise in extending the bandwidth of array processing [7], while Physics-informed Neural Networks (PINNs) incorporate acoustic principles to refine predictions [8]. DDPM has recently emerged as a powerful tool for sound field reconstruction, offering a probabilistic framework for generating accurate acoustic fields [9]. However, most of these approaches focus on specific frequency bands or parts of the RIR. A recent challenge focuses on generative models for synthesizing room acoustics as a data augmentation tool for speaker distance estimation tasks [10].

Our work explores the analogy between RIR reconstruction and image inpainting. By treating RIR matrices as images, we apply a diffusion model to reconstruct the full time span of RIRs. This novel approach enables robust and accurate RIR interpolation, achieving excellent performance in terms of NMSE and CD, even in scenarios where the microphones are sparsely distributed in the acoustic environment. The proposed method, supported by an experimental study using simulated acoustic environments, provides a strong foundation for potential real-world applications.

2. PROBLEM FORMULATION

This research aims to reconstruct RIRs for unmeasured locations using a limited number of measured RIRs. Given M measured RIRs in a room, the task is to estimate RIRs at L unmeasured

[*] ***Corresponding author**: sharon.gannot@biu.ac.il.*

locations, resulting in a total of $N = M + L$ locations. Each RIR is sampled at a frequency F_s and truncated to K samples, beyond which it falls into the noise floor.

This paper focuses on linear and semi-circular array configurations, as well as intermediate arc-shaped configurations, although the methodology can be extended to other setups. In this framework, we consider N microphone positions, of which only M randomly selected RIRs are measured, while the remaining L measurements are missing, as illustrated in Fig. 1 for a linear array.

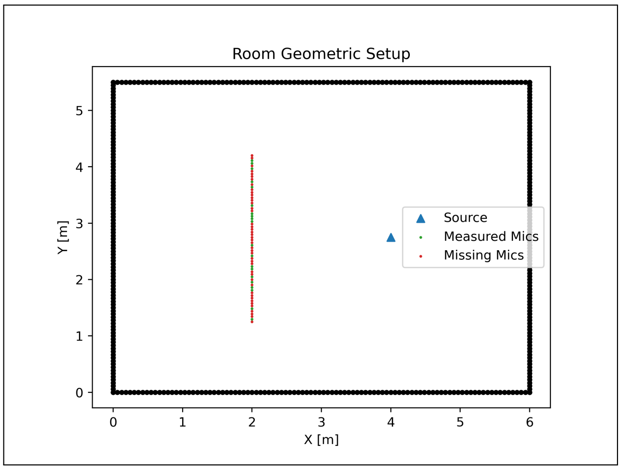

Figure 1. Geometric setup of a room with a source and a microphone array. Measured and missing microphones are marked in green and red dots, respectively. We aim to reconstruct the RIRs of the missing microphones.

Mathematically, let $\mathbf{H}$ be the matrix representing the RIRs, where $\mathbf{H} \in \mathbb{R}^{N \times K}$. We denote the available RIR measurements as $\mathbf{H}_{\text{measured}} \in \mathbb{R}^{M \times K}$. Our objective is to estimate the missing entries in $\mathbf{H}$ to obtain a complete matrix $\hat{\mathbf{H}} \in \mathbb{R}^{N \times K}$. Each column of $\mathbf{H}$, denoted $\mathbf{h}_i$, represents the RIR at the i-th location, where $1 \leq i \leq N$. Treating this matrix as an image, the problem is analogous to image inpainting, where the goal is to reconstruct the missing parts using the available data. Figure 2 shows a heatmap of the matrix $\mathbf{H}$, along with a zoomed-in view of one microphone's RIR. Our aim is to reconstruct the missing RIRs scattered throughout the array using the measured RIRs. Reconstructing missing RIRs requires leveraging the data's spatial and temporal structures. By addressing this challenge, we aim to develop a robust interpolation method that facilitates acoustic analysis and processing across various applications.

3. PROPOSED APPROACH

We formulate the problem of reconstructing missing RIRs as an image inpainting task. By representing the RIR data as an image, we can leverage the power of DDPMs to estimate the missing responses. Our work is inspired by previous research on image inpainting using diffusion models, notably [11], which demonstrated effective reconstruction of missing image regions using a pretrained diffusion model which was trained on the task of generating new images.

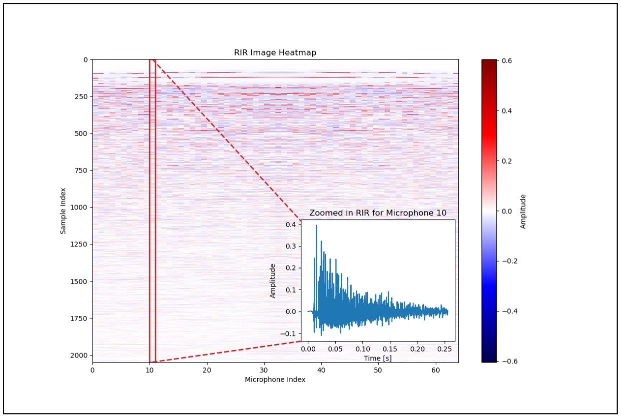

Figure 2. Heatmap of RIR matrix $\mathbf{H}$, with one zoomed-in RIR view. Each column in the matrix represents one microphone.

3.1 Inpainting with Diffusion Models

Lugmayr et al. [11] introduced RePaint, an inpainting method based on DDPMs. This approach utilizes a pre-trained model originally trained for general image generation. During inference, the model is adapted to the inpainting task by conditioning it on the known parts of an image while generating new content for the missing regions. At each diffusion step, the model is guided to remain consistent with the observed parts, ensuring that only the missing regions are reconstructed while the known areas are preserved. This method allows for flexible inpainting without requiring prior knowledge of the mask pattern.

This iterative refinement aligns well with our problem, where missing RIR data should be reconstructed to closely resemble the original responses without prior knowledge of the missing microphone positions.

We adopt OpenAI's DDPM architecture [1] with necessary modifications to accommodate RIR-matrix images. While the original model is designed for natural images, RIR data exhibits distinct statistical properties. Training the model on a dedicated small RIR dataset allows it to capture these characteristics, leading to more accurate reconstructions.

During inference, a masked RIR image is fed into the trained diffusion model, which iteratively reconstructs the missing regions. The output is a complete RIR image. Finally, the recon-

[1] https://github.com/openai/guided-diffusion

structed image is converted to its original matrix form by transforming grayscale pixel values into response amplitudes. Only the newly inpainted regions are retained, representing the reconstructed RIR.

3.2 Image Representation of the RIR Set

To apply inpainting techniques, we recast the RIR data into an image-like format. Given an array configuration, we arrange the RIRs into a 2D matrix where each column corresponds to an RIR of length K from a specific microphone position. Different numbers of missing microphones and RIR lengths can also be accommodated. This will result in images of varying width and height dimensions. The resulting matrix is treated as a grayscale image, with intensity values representing normalized RIR amplitudes. This format enables structured processing while retaining spatial and temporal information.

Since DDPMs are typically trained on fixed-size images, we split the RIR matrix into patches of 64×64 pixels, corresponding to 64 possible microphone positions and 64 RIR taps. If the length of the RIR exceeds 64, as is often the case, we divide the image into multiple patches, each representing a different portion of the RIR.

To address the issue of lower reconstruction quality at the edges of the patches due to the lack of surrounding context, we introduce an overlap of 25% between adjacent patches. We also normalize each patch to the range -1 to 1, allowing the network to reconstruct each patch independently of the energy level of that part of the response. After reconstruction, these patches are reassembled into a complete image by rescaling each patch to its original energy, discarding the overlapping regions, and retaining only the central portions of the patches. This approach balances computational efficiency and reconstruction accuracy and ensures a seamless reconstruction by eliminating duplicates and maintaining continuity.

In cases where the microphone configuration has fewer than 64 microphones, we pad the image with duplicated columns to ensure an image width of 64 pixels. This preserves the model's expected input dimensions while minimizing distortions in the reconstruction process.

To simulate missing measurements, we generate masks of varying percentages by zeroing out randomly selected columns in the RIR image. These masks represent the unmeasured microphone locations. The masked image, along with its corresponding mask, are then fed as input to the diffusion model.

4. EXPERIMENTS

In this section, we describe our experiments using artificial RIRs generated by the Pyroomacoustics package. [2]

4.1 Experiment Setup

The simulated database of RIRs comprises multiple microphone array configurations: a uniform linear array (ULA), a semi-circular array, and intermediate arcs. The ULA configuration uses 64 microphones and spans a length of 3 meters, resulting in a 4 cm distance between adjacent microphones. The semi-circular array configuration uses 64 microphones with a 1.5-meter radius, resulting in a 7.3 cm distance between adjacent microphones. The simulated room dimensions are $6 \times 5.5 \times 2.8\,\mathrm{m}$ (length, width, and height, respectively). The simulation uses a sampling frequency of $F_s = 8\,\mathrm{kHz}$.

The data is split into training and inference sets. The training set consists of 176 randomly selected patches from 8 RIR images, corresponding to various microphone array configurations. The source positions were randomly selected from 9 positions on a semi-circle with a radius of 2 meters from the array's center, as depicted in Fig. 14. During training, the reverberation time (T_{60}) was fixed at 0.3 seconds across all frequency bands. Each RIR was truncated to 1024 samples, corresponding to a duration of 0.128 seconds.

As previously mentioned, during training, the model learns to generate new images from the distribution of the training dataset. During inference, the model generates new images while conditioning on the known parts of the image, which correspond to the measured responses.

To introduce variability in absorption coefficients across frequency bands, we selected "smooth brickwork 10 mm pointing," from the Pyroomacoustics material database as the wall material for the inference dataset. Using this material and the specified room dimensions corresponds to a full-band reverberation time (T_{60}) of 0.6 seconds. The evaluation was carried out using a ULA, a semi-circular array, and intermediate arc configurations. All nine different source positions were tested. Each RIR was truncated to 2048 samples, corresponding to a duration of 0.256 seconds. In each trial, we randomly removed a percentage of the microphone measurements, varying the ratio of missing microphones from 10% to 90%.

4.2 Performance Measures and a Baseline Method

The results of our experiments are analyzed using two quality measures, comparing the estimated and the ground truth RIR for the M missing microphones. The first is the Normalized Mean Square Error (NMSE) in dB, defined as (see [5]):

$$\mathrm{NMSE}(\mathbf{H}, \hat{\mathbf{H}}) = 10 \log_{10} \left(\frac{1}{M} \sum_{i=1}^{M} \frac{\|\hat{\mathbf{h}}_i - \mathbf{h}_i\|^2}{\|\mathbf{h}_i\|^2} \right), \qquad (1)$$

where $\hat{\mathbf{h}}_i \in \mathbb{R}^{N \times 1}$ is the estimate of the ith RIR corresponding to the ith column of $\hat{\mathbf{H}}$. The second is the Cosine Distance (CD), defined as (see also [12]):

$$\mathrm{CD}(\mathbf{H}, \hat{\mathbf{H}}) = \frac{1}{M} \sum_{i=1}^{M} \left(1 - \left(\frac{\mathbf{h}_i^\top \hat{\mathbf{h}}_i}{\|\mathbf{h}_i\| \|\hat{\mathbf{h}}_i\|} \right)^2 \right). \qquad (2)$$

[2] https://pyroomacoustics.readthedocs.io/en/pypi-release/

The value $\mathrm{CD}(\mathbf{H}, \hat{\mathbf{H}}) = 1$ is obtained if all estimates are orthogonal to the corresponding true RIR for all M missing values (i.e., all estimates are the worst possible), and $\mathrm{CD}(\mathbf{H}, \hat{\mathbf{H}}) = 0$ if all estimated and true RIRs are perfectly aligned. The CD is particularly useful in audio applications [12].

Finally, we used the Spline Cubic Interpolation (SCI) technique as a baseline method [13].

4.3 Results

In this section, we present and analyze the performance of the proposed method and compare it with the baseline method.

Figures 3 and 4 depict the NMSE and CD, respectively, for different mask ratios for linear array configuration as shown in Fig. 1. Our method improves the NMSE by 3 to 7 dB and the CD measure over the baseline by approximately 0.3, depending on the mask ratio.

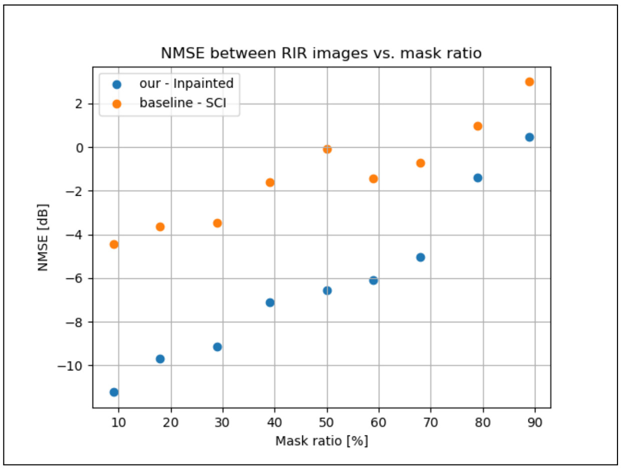

Figure 3. NMSE for inpainted and baseline SCI vs. mask ratio.

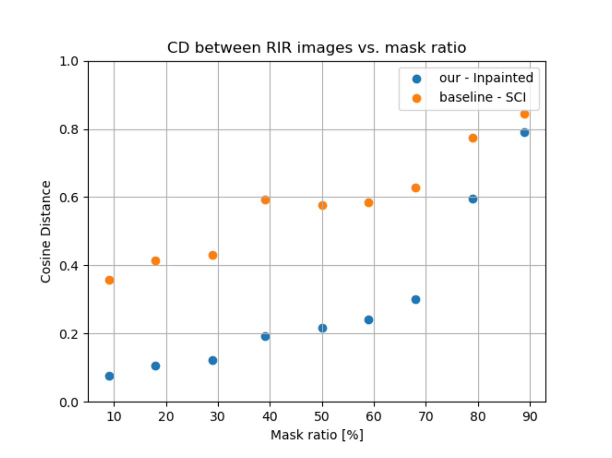

Figure 4. CD for inpainted and baseline SCI vs. mask ratio.

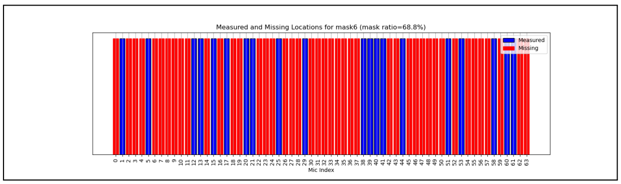

Figure 5. Linear array configuration: Measured (blue) and missing (red) microphones.

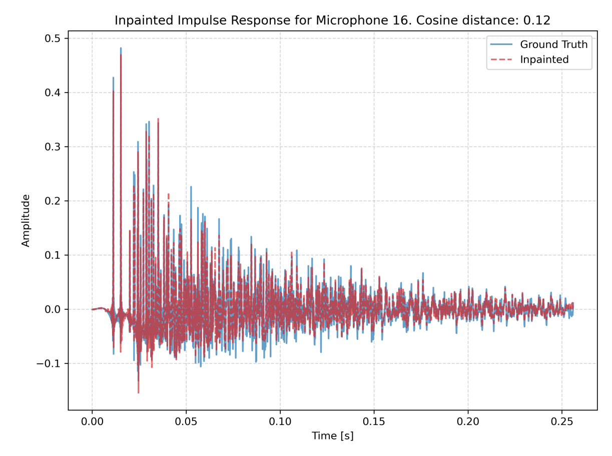

Figure 6. Reconstructed and ground truth impulse response for microphone No. #16.

We now focus on the reconstruction results for a 70% mask ratio. Figure 5 provides a detailed view of the microphone indices, describing which microphone signals are measured (blue) and which are missing (red). In Figs. 6 and 7, we present the reconstructed and ground truth RIRs for microphones #16 and #48, respectively. Microphone #16 is located very close to the measured microphones, while microphone #48 is situated in a region with sparse measurements, leading to better reconstruction for the former. Yet, even in the more challenging case of microphone #48, the reconstructed RIR successfully captures the main features, including both the direct and early arrivals. The CD for microphone #48 is relatively low at 0.37 (but higher than CD $= 0.12$ for microphone #16), demonstrating the model's ability to accurately infer and reconstruct acoustic reflections even in areas with limited measured data. In Fig. 8, we further analyze the acoustic properties of RIR #48 by examining its Energy Decay Curve (EDC). It is evident that the EDC of the reconstructed RIR closely resembles that of the ground truth. Moreover, the estimated full-band reverberation time de-

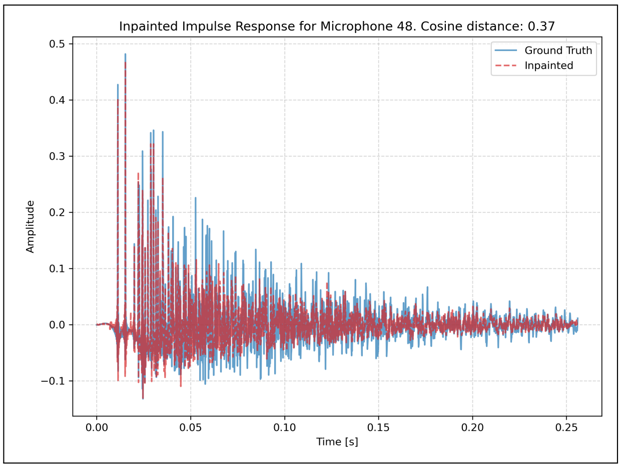

Figure 7. Reconstructed and ground truth impulse response for microphone No. #48.

rived from the EDC slope, $T_{60} = 0.64$ seconds, closely matches the ground truth value of $T_{60} = 0.6$ seconds. Despite the large percentage of missing microphones, our method demonstrates favorable performance for the ULA configuration, generating a reconstructed RIRs that closely align with the ground truth responses.

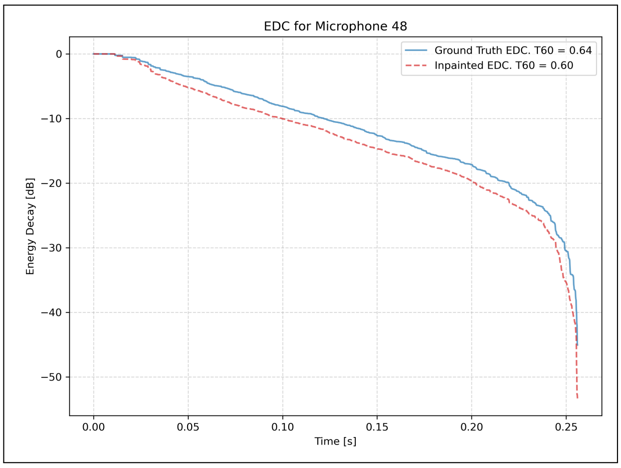

Figure 8. EDC of reconstructed and ground truth impulse response for microphone No. #48, and the corresponding T_{60}.

Next, we evaluate the performance of the proposed method for the semi-circular microphone array. First, Fig. 9 illustrates the room and several array curvatures. We begin by examining the semi-circular array labeled as array #10 in the figure.

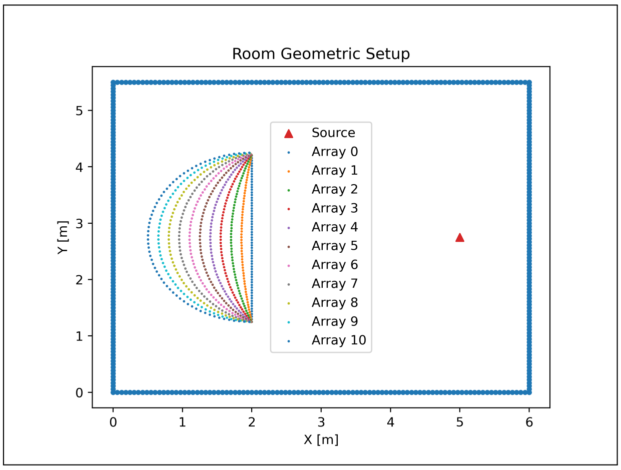

Figure 9. Room with several array curvatures from a linear to a semi-circular configuration.

The reconstruction results for the semi-circular array, as measured by the CD metric, are presented in Fig. 10. Our method significantly outperforms the baseline interpolation approach, with improvements of 0.2–0.6 in the CD measure, up to a mask ratio of 70%.

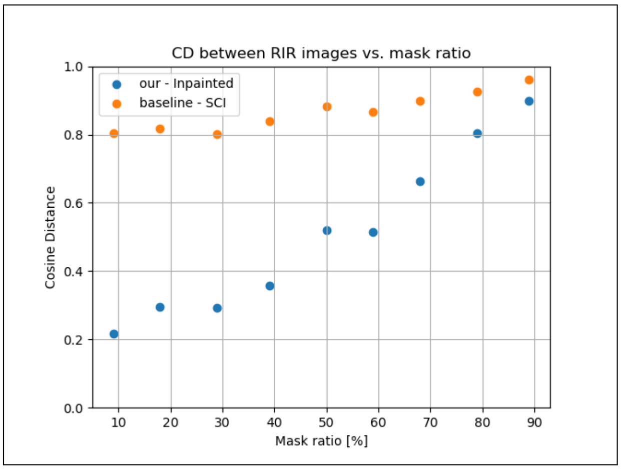

Figure 10. CD for semi-circular array vs. mask ratio.

Despite these improvements, the inpainting algorithm's performance for the semi-circular array is inferior compared to the linear array. This difference can be attributed to the geometric challenges posed by the curved reflection patterns in the semi-circular array, which are more complex to reconstruct than the straight-line patterns found in the linear array.

The influence of array curvature on performance is further explored in Figs. 11 and 12, which present the NMSE and CD measures, respectively. As the array curvature increases, performance degradation becomes evident, suggesting that the inpainting task is more manageable when the reflection patterns are straight rather than curved.

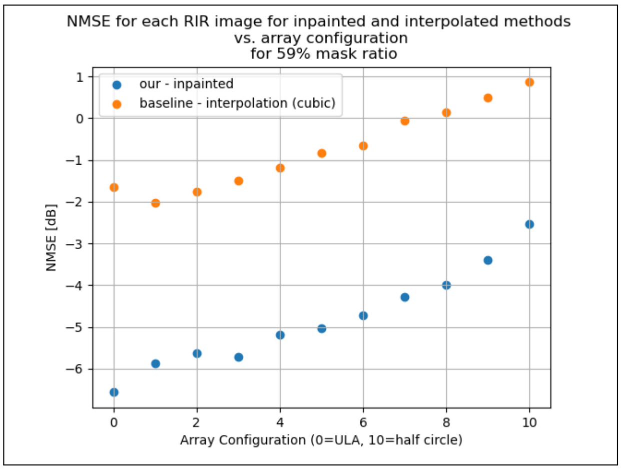

Figure 11. NMSE for different array curvatures.

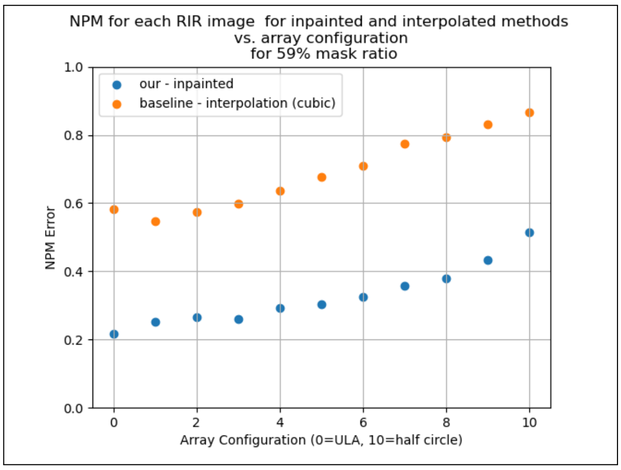

Figure 12. CD for different array curvatures.

The RIR images in grayscale for the linear array, the semi-circular array, and two intermediate configurations are presented in Fig. 13. These images highlight differences in reflection patterns: the linear array exhibits straight-line reflections that are easier to reconstruct, whereas the semi-circular array produces curved reflections, which pose greater challenges during inpainting. These findings emphasize that while the model adapts well to semi-circular arrays, it achieves superior performance when the reflection paths are straight.

Finally, in Fig. 14, we show the room setup with 9 loudspeaker angles. Figure 15 demonstrates that the best results are obtained for sources located at 90° ('broadside') relative to the array, while higher errors are observed for sources positioned at 10° or 170° (towards 'endfire').

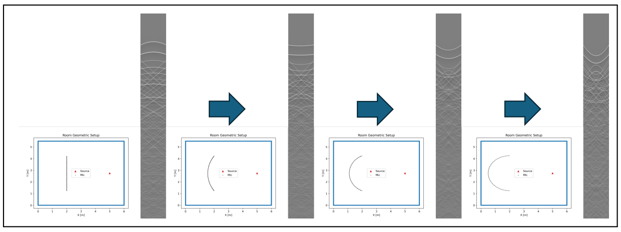

Figure 13. RIR images in grayscale for the linear array, circular array, and intermediate configurations.

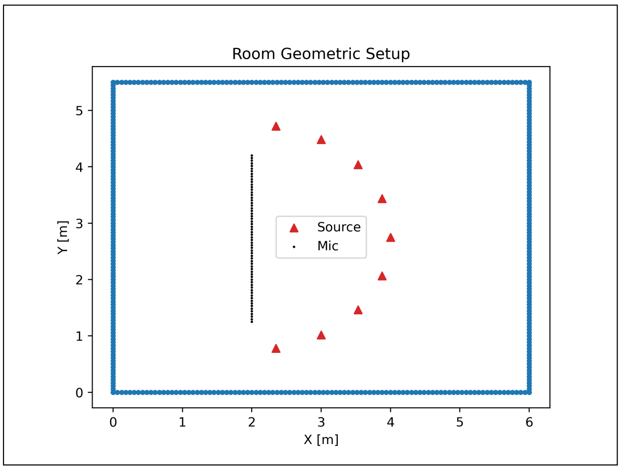

Figure 14. Geometric setup of the room with different source angles.

To further illustrate this, Fig. 16 presents several grayscale images of RIRs, ranging from broadside to endfire configurations. When the source is positioned in front of the array, i.e., in the broadside configuration, the image exhibits greater symmetry, making it easier to inpaint and reconstruct the missing points. However, when the source is located at endfire angles, it becomes more challenging to complete the impulse response for the microphones on the opposite side, which are farther away. Additionally, the lines in the endfire configuration are sharper, while those from the broadside configuration are smoother.

5. DISCUSSION

We addressed the challenge of acquiring RIR measurements, which are essential for characterizing a room's acoustic properties but are resource-intensive to collect. We propose leveraging super-resolution techniques, traditionally used in imaging,

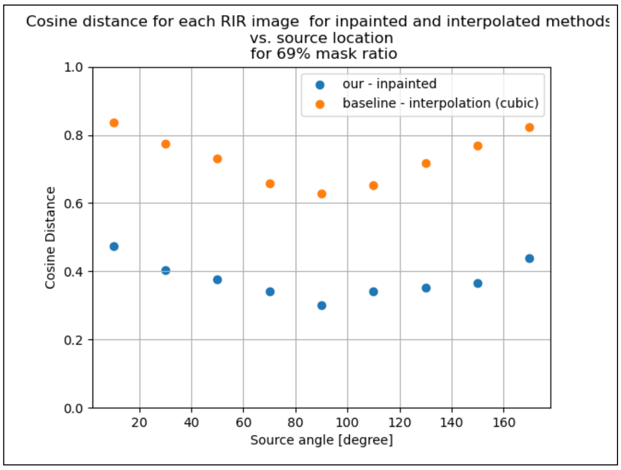

Figure 15. CD for inpainted and baseline SCI for different source location.

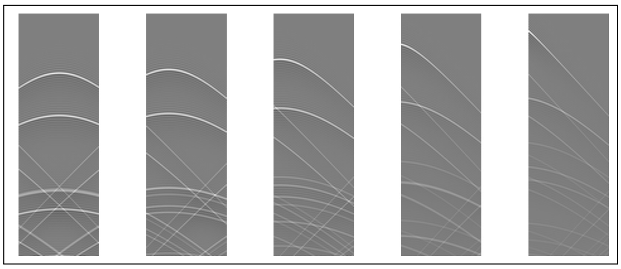

Figure 16. Comparison of RIR images for different angles: from 90° - broadside (on the left) to 10° - endfire (on the right).

to interpolate or predict RIRs at unmeasured locations within a room. This method utilizes existing RIR data to generate high-resolution acoustic mappings without the need for exhaustive measurements, enabling applications in sound source localization, separation, and augmented reality.

Our simulation results show that the proposed method generalizes effectively beyond the trained configurations, allowing the generation of RIRs for different microphone arrays and even for rooms that were not part of the training set. Although tested with simulated RIRs, we believe that this research opens the door to generating additional data from limited real-world measurements.

6. REFERENCES

[1] O. Thiergart, G. Del Galdo, M. Taseska, and E. A. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2583–2594, 2013.

[2] M. Pezzoli, F. Borra, F. Antonacci, S. Tubaro, and A. Sarti, "A parametric approach to virtual miking for sources of arbitrary directivity," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2333–2348, 2020.

[3] M. Pezzoli, M. Cobos, F. Antonacci, and A. Sarti, "Sparsity-based sound field separation in the spherical harmonics domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1051–1055, 2022.

[4] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "Sound field separation in a mixed acoustic environment using a sparse array of higher order spherical microphones," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pp. 151–155, 2017.

[5] E. Zea, "Compressed sensing of impulse responses in rooms of unknown properties and contents," *Journal of Sound and Vibration*, vol. 459, p. 114871, 2019.

[6] M. Pezzoli, D. Perini, A. Bernardini, F. Borra, F. Antonacci, and A. Sarti, "Deep prior approach for room impulse response reconstruction," *Sensors*, vol. 22, no. 7, p. 2710, 2022.

[7] E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstoft, "Generative models for sound field reconstruction," *The Journal of the Acoustical Society of America (JASA)*, vol. 153, no. 2, pp. 1179–1190, 2023.

[8] M. Olivieri, M. Pezzoli, F. Antonacci, and A. Sarti, "A physics-informed neural network approach for nearfield acoustic holography," *Sensors*, vol. 21, no. 23, p. 7834, 2021.

[9] F. Miotello, L. Comanducci, M. Pezzoli, A. Bernardini, F. Antonacci, and A. Sarti, "Reconstruction of sound field through diffusion models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1476–1480, 2024.

[10] J. Lin, G. Götz, H. S. Llopis, H. Hafsteinsson, S. Guðjónsson, D. G. Nielsen, F. Pind, P. Smaragdis, D. Manocha, J. Hershey, T. Kristjansson, and M. Kim, "Generative data augmentation challenge: Synthesis of room acoustics for speaker distance estimation," *arXiv preprint arXiv:2501.13250*, 2025.

[11] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, 2022.

[12] D. R. Morgan, J. Benesty, and M. M. Sondhi, "On the evaluation of estimated impulse responses," *IEEE Signal Processing Letters*, vol. 5, no. 7, pp. 174–176, 1998.

[13] C. De Boor, *A practical guide to splines*, vol. 27. springer New York, 1978.