FEW-SHOT ROOM IMPULSE RESPONSE INTERPOLATION IN LATENT DOMAINS

Jackie Lin[1,2*], *Yoshiki Masuyama*[1], *Christoph Boeddeker*[1], *Julius Richter*[1], *Gordon Wichern*[1], *Minje Kim*[2], *Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, USA,
[2]University of Illinois Urbana-Champaign, USA

ABSTRACT

We address the problem of few-shot room impulse response (RIR) interpolation, where the goal is to estimate the RIR at an arbitrary receiver location given a small set of location-labeled reference RIRs from that same room. Naïve signal-domain methods, such as direct interpolation in the time domain, often produce inaccurate results. To overcome this, we leverage an audio variational autoencoder (VAE) to map RIRs into a compact and smooth latent space that is amenable to interpolation. We explore both training-free and network-based approaches in this latent domain, including distance-based kernel regression and geometry-informed attention pooling. Our results show that simple kernel methods serve as strong baselines, while lightweight neural networks further improve performance. We present a promising, flexible latent-domain framework for few-shot RIR interpolation and provide insight into the interpolation behavior of latent representations across diverse acoustic conditions.

Index Terms— Room impulse response (RIR), few-shot interpolation, latent domain, variational autoencoder (VAE)

1. INTRODUCTION

Room impulse responses (RIRs) describe the acoustic transfer path between a sound source and a receiver, including the direct sound, early reflections, and late reverberation [1]. They are fundamental to spatial audio rendering, augmented and virtual reality, dereverberation, and source localization [2, 3]. In many applications, however, the source or receiver moves continuously while RIR measurements are available only at a finite number of positions. Moreover, attempts to dynamically simulate RIRs especially using high-fidelity wave-based simulators are computationally expensive. This motivates *RIR interpolation*: estimating the RIR at an unmeasured query location from a sparse set of reference RIRs [4–7].

A common use case is dynamic acoustic rendering, where receiver motion should result in smooth and physically plausible acoustic changes. Simple time-domain interpolation, analogous in spirit to panning [8], is computationally efficient, but directly averaging waveforms often produces unrealistic results. Small geometric changes shift reflection arrivals, causing time-domain interpolation to smear reflection peaks and degrade perceptual quality. These limitations motivate interpolation on representations that better capture the underlying acoustic structure rather than on raw waveforms.

More recently, optimal-transport-based methods match reflection structures rather than directly averaging samples [9–11]. Meanwhile, physics-informed methods design kernel functions based on the Helmholtz equation and perform kernel interpolation [6, 12]. These approaches improve over naïve interpolation by leveraging physical models of sound propagation.

Learning-based methods have also gained increasing attention due to their powerful modeling capability and flexibility [13–20]. Neural fields model an RIR as a continuous function over spatial coordinates, enabling RIR prediction at arbitrary positions after room-specific training [21, 22]. Other neural approaches typically use multimodal cues [23] or differentiable room rendering [17, 18] to estimate responses at novel locations. These methods demonstrate the promise of machine learning for spatially continuous acoustic modeling, but may require room-specific optimization, dense measurements, and/or visual cues. A direction more relevant to ours is *few-shot cross-room RIR prediction*, where the goal is to estimate an RIR in an unseen room from only a handful of reference measurements [24, 25]. For example, xRIR combines acoustic features from a few reference RIRs with visual cues to interpolate RIRs in new environments [25]. However, they have focused on interpolating STFT magnitudes of RIRs and have relied on separate phase reconstruction without considering spatial relationships.

In this work, we study *few-shot RIR interpolation in latent domains*. Rather than interpolating directly in the time or STFT domain, we train a variational autoencoder (VAE) to map RIRs to a compact latent representation, where latent RIR interpolation is performed before decoding back to the time domain. We expect that the trained latent space is smoother and more structured than raw signal space, making it more suitable for interpolation. To investigate this hypothesis, we assess the performance of *training-free kernel interpolation* which is a simple weighted average in the VAE latent space [26, 27], similar to an approach taken by [20].

We then examine whether performance can be further improved using a *learning-based latent interpolation method* inspired by re-bottlenecking [28]. Specifically, a lightweight network predicts the target RIR latent from reference latents and geometry, which can learn complex geometric relationships beyond fixed distance kernels. The proposed framework enables flexible few-shot RIR interpolation across diverse rooms, with experiments revealing that the learned latent space supports training-free interpolation surprisingly effectively.

2. PROBLEM DEFINITION

We consider the problem of predicting an RIR from a fixed source location at a queried receiver position $\boldsymbol{p}_q \in \mathbb{R}^3$, given K reference RIRs $\boldsymbol{H}_{\text{ref}} = \{\boldsymbol{h}_i\}_{i=1}^K$, where $\boldsymbol{h}_i \in \mathbb{R}^N$. The RIRs in $\boldsymbol{H}_{\text{ref}}$ share the same source location but correspond to K different receiver positions $\boldsymbol{P}_{\text{ref}} = \{\boldsymbol{p}_i\}_{i=1}^K$ in the same room. A predictor $\mathcal{P}$ estimates the RIR $\boldsymbol{h}_q \in \mathbb{R}^N$ at the desired position $\boldsymbol{p}_q$ as

$$\hat{\boldsymbol{h}}_q = \mathcal{P}\left(\boldsymbol{H}_{\text{ref}}, \boldsymbol{P}_{\text{ref}}, \boldsymbol{p}_q\right). \tag{1}$$

*This work was done during an internship at MERL.

When reference receivers are sufficiently close to the query location, nearest neighbor selection can approximate the target RIR well. However, when the query is far from the references, accurate estimation requires a more sophisticated model capable of generalizing across diverse acoustic conditions. Hence, our method is based on few-shot cross-room RIR prediction. In contrast to existing learning-based methods [21, 22], we do not require room-specific optimization with a large number of reference measurements.

3. PROPOSED METHOD

3.1. Overview

We investigate whether a generic latent domain learned for RIR reconstruction is sufficiently compact and locally smooth to support few-shot interpolation. Our proposed method uses a VAE $\mathcal{F}$, which comprises an encoder $\mathcal{F}_{\text{enc}}$ and decoder $\mathcal{F}_{\text{dec}}$ trained to compress and reconstruct RIRs. We adapt a time-domain convolutional VAE architecture used in neural audio coding models [29].

We then study two strategies for estimating the latent representation of the queried RIR from the encoded reference RIRs. First, we explore a weighted average of encoded reference RIRs using kernel-based weighting schemes dependent on the physical distance between the reference and query coordinates. Second, we borrow the concept of "Re-bottlenecking" [28] to learn interpolation in the VAE latent space, via a small additional network, to predict the queried RIR's latent representation more efficiently.

Our method is related to [20], which performs RIR interpolation in the latent space of an encoder-decoder. This former work focuses on training and evaluation on RIRs from a unique room, and the encoder-decoder is optimized specifically for that room. On the other hand, we explore few-shot cross-room RIR prediction, where the VAE is pre-trained on RIRs from diverse rooms and frozen.

3.2. Distance-based weighted average

We investigate a training-free weighted average approach to predict the queried RIRs in the VAE latent space. First, the reference RIRs $\boldsymbol{H}_{\text{ref}}$ are encoded by the VAE into reference latents $\boldsymbol{Z}_{\text{ref}} = \{\boldsymbol{z}_i\}_{i=1}^{K}$, where $\boldsymbol{z}_i = \mathcal{F}_{\text{enc}}(\boldsymbol{h}_i) \in \mathbb{R}^{L\times T}$, L denoting the latent dimension and T the number of time frames. Then, the prediction is realized by a weighted sum of these latents, where the weight for each reference latent $\boldsymbol{z}_i$ is calculated based on reference location $\boldsymbol{p}_i$ and queried receiver location $\boldsymbol{p}_q$.

Specifically, the weight is formulated based on the Nadaraya–Watson (NW) estimator [26, 27]:

$$w_i = \frac{\mathcal{K}(\boldsymbol{p}_i, \boldsymbol{p}_q)}{\sum_{j=1}^{K} \mathcal{K}(\boldsymbol{p}_j, \boldsymbol{p}_q)}, \tag{2}$$

where $\mathcal{K}(\boldsymbol{p}_i, \boldsymbol{p}_q) \geq 0$ is a function[1] taking a large value when $\boldsymbol{p}_q$ is close to $\boldsymbol{p}_i$. A prediction of the queried RIR latent is then obtained as the weighted sum of the reference latents:

$$\bar{\boldsymbol{z}}_w = \sum_{i=1}^{K} w_i \boldsymbol{z}_i. \tag{3}$$

The predicted RIR latent is then converted back to the time domain using the decoder, i.e., $\hat{\boldsymbol{h}}_q = \mathcal{F}_{\text{dec}}(\bar{\boldsymbol{z}}_w)$. The interpolation performance depends on the weighting kernel. We compare inverse-distance (ID) weighting and radial-basis-function (RBF) kernels in the NW estimator as summarized in Table 1. Furthermore, we include direct ID weighting in the time domain, i.e., $\bar{\boldsymbol{h}}_w = \sum_i^K w_i \boldsymbol{h}_i$ to provide a direct comparison with latent-domain interpolation.

[1]For simplicity, we refer to the inverse-distance function in Table 1 as a kernel function, although it is singular at zero distance.

Table 1: Distance-based kernel functions, where the distance is the Euclidean distance between reference receiver position and queried receiver position, i.e., $d_i = \|\boldsymbol{p}_i - \boldsymbol{p}_q\|_2$.

Kernel	**Equation**
RBF kernel (bandwidth σ)	$\mathcal{K}(\boldsymbol{p}_i, \boldsymbol{p}_q) = \exp\left(-\frac{d_i^2}{2\sigma^2}\right)$
Inverse-distance (ID) kernel	$\mathcal{K}(\boldsymbol{p}_i, \boldsymbol{p}_q) = \frac{1}{d_i}$

3.3. Geometry-informed latent attention pooling

While the distance-based weighted average provides a strong training-free baseline, fixed weighting schemes based only on pairwise distances may not capture more complex geometric relationships between reference and query locations. We thus investigate a lightweight network operating directly in the frozen VAE latent space, inspired by the concept of re-bottlenecking [28].

The proposed model $\mathcal{G}$ predicts the latent representation $\boldsymbol{z}_q$ corresponding to the queried RIR using the reference latents, their receiver coordinates, and the queried location:

$$\hat{\boldsymbol{z}}_q = \mathcal{G}\big(\boldsymbol{Z}_{\text{ref}}, \boldsymbol{P}_{\text{ref}}, \boldsymbol{p}_q\big). \tag{4}$$

Rather than directly predicting the queried latent, the model uses attention to first compute a geometry-aware interpolation baseline $\boldsymbol{z}_{\text{base}}$ and then refine it by predicting $\Delta\boldsymbol{z}$. This combines the stability of explicit interpolation with the flexibility of learned refinement.

For each reference i, we construct a reference token $\boldsymbol{r}_i$ that fuses the reference's latent representation with geometric relationship between i and the queried position. To encode the geometry information, we compute relative geometric features $\Delta_i = \boldsymbol{p}_i - \boldsymbol{p}_q$ and $d_i = \|\Delta_i\|_2$, and form a concatenated geometry vector $\boldsymbol{g}_i = [\Delta_i, d_i]$. A Fourier feature mapping $\tilde{\boldsymbol{g}}_i = \Phi(\boldsymbol{g}_i)$ is applied to encode higher-frequency geometric variation [30]. Each reference latent $\boldsymbol{z}_i$ is averaged over time to obtain a summary representation $\bar{\boldsymbol{z}}_i \in \mathbb{R}^L$, which is concatenated with the geometric embedding $\tilde{\boldsymbol{g}}_i$ and projected into a reference token $\boldsymbol{r}_i$ via a multilayer perceptron (MLP). Separately, the query position $\boldsymbol{p}_q$ is projected to a query token $\boldsymbol{q}$.

Attention scores are computed using a query-conditioned scoring MLP $\mathcal{U}_{\text{score}}$ followed by softmax normalization:

$$s_i = \mathcal{U}_{\text{score}}([\boldsymbol{q}, \boldsymbol{r}_i, \tilde{\boldsymbol{g}}_i]), \quad \alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{K} \exp(s_j)}. \tag{5}$$

These weights define a geometry-aware interpolation baseline $\boldsymbol{z}_{\text{base}} = \sum_{i=1}^{K} \alpha_i \boldsymbol{z}_i$. To refine this estimate, we compute a local context vector by applying the same attention weights to projected reference tokens, $\boldsymbol{c}_l = \sum_{i=1}^{K} \alpha_i \boldsymbol{W}_v \boldsymbol{r}_i$, and a global context vector by averaging over the reference set, $\boldsymbol{c}_g = \boldsymbol{W}_g\big(\frac{1}{K}\sum_{i=1}^{K} \boldsymbol{r}_i\big)$, where $\boldsymbol{W}_v$ and $\boldsymbol{W}_g$ are learned projections. The local context $\boldsymbol{c}_l$, global context $\boldsymbol{c}_g$, and query token $\boldsymbol{q}$ are concatenated to form a conditioning vector, which is broadcast along the temporal dimension and passed with $\boldsymbol{z}_{\text{base}}$ to a lightweight convolutional network. This network predicts a residual correction $\Delta\boldsymbol{z}$, from which the final prediction is obtained as

$$\hat{\boldsymbol{z}}_q = \boldsymbol{z}_{\text{base}} + \Delta\boldsymbol{z}. \tag{6}$$

We refer to this model as geometry-informed latent attention pooling (GLAP). GLAP remains permutation-invariant with respect to the reference set, supports a variable number of references, and enables learned geometry-aware latent interpolation beyond fixed kernel weighting. In our experiments, the model is trained with a reference set size $K = 4$.

4. EXPERIMENTAL SETUP

RIR Variational Autoencoder: We first train a VAE for RIR reconstruction, and use its latent space as the representation domain for all subsequent experiments. The architecture is based on the Stable Audio Open 2.0 VAE [29] which consists of convolutional downsampling and upsampling blocks. The VAE design compresses 22.05 kHz audio to 64-channel latents at 21.5 Hz. The VAE is trained from scratch using a combination of multiresolution STFT (MRSTFT), KL divergence, adversarial, and feature matching losses on the simulated RIR dataset described below.

GLAP: With the VAE frozen, the RIR training set is encoded into latent representations using $\mathcal{F}_{\text{enc}}$. GLAP is then trained using a weighted combination of latent L2 and decoded MRSTFT losses, with gradients propagated through the frozen decoder.

Training Dataset: The training dataset was generated using Pyroomacoustic's hybrid image source and ray tracing RIR engine [31]. The simulated dataset consists of 165k RIRs from over 2500 shoebox rooms with randomized absorption coefficients and room dimensions. In each room, two sources and a 3×11 rectangular microphone array randomly rotated in the horizontal plane are placed randomly. The grid spacing for the microphones in the array is 0.1 m. The source-receiver distance range is 0.204-7.27 m, the mean 2.71 m, and the median 2.22 m. The RIRs are cropped to 0.743 s, as later arrivals generally offer diminishing returns in information.

Test Dataset: The test dataset consists of 32k RIRs simulated with the same parameters as the training set but using a different random seed to ensure unseen rooms and source-receiver configurations. For each test sample, the reference RIR locations are constrained to form a convex hull containing the query location, limiting the problem to interpolation and enabling stable interpretation of the kernel weights.

Evaluation Metrics: We report RIR interpolation errors in the latent domain as well as in the signal domain, namely, MRSTFT error, energy decay curve mean absolute error (EDC [dB]), reverberation time 60 dB percentage absolute error (RT [%]), early decay time absolute error (EDT [s]), clarity absolute error (C50 [dB]), and direct to reverberant ratio absolute error (DRR [dB]).

Baselines: Several naïve baselines are reported, namely random selection from reference set $\boldsymbol{h}_{i_{\text{rand}}}$, nearest neighbor $\boldsymbol{h}_{i_{\text{NN}}}$, weighted average in the signal domain $\bar{\boldsymbol{h}}_w$ (weights based on inverse distance), and simple average in the latent domain $\bar{\boldsymbol{z}}$.

5. RESULTS AND DISCUSSION

The second and third sections of Table 2 show the overall performance of the naïve baselines and our latent interpolation methods, respectively. Before comparing interpolation methods, we first establish the representational limit imposed by the frozen VAE. As shown in Table 2 row $\boldsymbol{z}_q$, reconstructing the ground-truth queried RIR through the encoder-decoder incurs non-negligible error, setting a reconstruction floor shared by all latent-domain interpolation methods. Consequently, the following results evaluate interpolation within the representational capacity of the learned latent space.

Table 2: Errors of RIR estimates (**"Est."**) on the test split for convex hull reference sets of size $K = 4$. Column **L2** is the latent-domain MSE between the ground-truth queried latent $\boldsymbol{z}_q = \mathcal{F}_{\text{enc}}(\boldsymbol{h}_q)$ and **Est.** (or its latent representation for a time domain signal). The remaining columns are signal-domain errors between the ground-truth queried RIR and **Est.** (decoded by $\mathcal{F}_{\text{dec}}$ if **Est.** is a latent). The three sections of rows correspond to VAE reconstruction, naïve time-domain interpolation, and latent-domain interpolation methods.

Method	Est.	L2	MRSTFT	EDC	RT [%]	EDT	C50	DRR
	$\boldsymbol{z}_q$	–	0.691	2.76	4.04	0.114	1.06	2.20
Rand. N.	$\boldsymbol{h}_{i_{\text{rand}}}$	0.454	0.937	0.85	3.41	0.025	0.75	2.59
Near. N.	$\boldsymbol{h}_{i_{\text{NN}}}$	0.309	0.842	0.30	2.80	0.018	0.46	1.24
Near. N.	$\boldsymbol{z}_{i_{\text{NN}}}$	0.309	0.904	2.91	5.00	0.035	1.11	2.52
ID	$\bar{\boldsymbol{h}}_w$	0.683	1.283	49.00	6.10	0.031	0.91	4.81
Mean N.	$\bar{\boldsymbol{z}}$	0.248	0.936	3.77	5.98	0.035	1.09	2.87
ID	$\bar{\boldsymbol{z}}_w$	0.220	0.902	2.79	5.11	0.036	1.02	2.49
$\text{RBF}_{0.075}$	$\bar{\boldsymbol{z}}_w$	0.280	0.901	2.84	5.04	0.036	1.07	2.48
$\text{RBF}_{0.1}$	$\bar{\boldsymbol{z}}_w$	0.262	0.898	2.80	4.92	*0.034*	1.06	2.44
$\text{RBF}_{0.2}$	$\bar{\boldsymbol{z}}_w$	0.237	0.898	2.74	5.04	0.035	1.02	**2.43**
GLAP	$\hat{\boldsymbol{z}}_q$	**0.217**	**0.897**	**1.77**	**4.44**	*0.034*	**0.68**	2.49

5.1. Distance-based kernel weighting results

Figure 1 disaggregates the averages reported in Table 2 by room size (small, medium, large) and material absorptivity (anechoic, high, medium, low), further grouping the results by the distance between the query receiver and its nearest reference. The nearest neighbor baseline $\boldsymbol{h}_{i\text{NN}}$ is competitive when the closest reference is sufficiently near the query, typically in reflective rooms, as shown in Fig. 1 rows (b), (c), and (d). As the nearest reference distance increases, the $\boldsymbol{h}_{i\text{NN}}$ error grows more rapidly and in some cases exceeds that of all explored latent methods, as in Fig. 1 (d.i) and (b.iii). In anechoic conditions, as in Fig. 1 row (a), latent interpolation performs comparably or slightly better, indicating that local interpolation among latent representations of anechoic RIRs produces correspondingly smooth shifts in the decoded direct-path arrival.

Table 2 highlights the importance of the operation domain. The inverse-distance average in the time domain performs $\bar{\boldsymbol{h}}_w$ substantially worse than selecting the nearest neighbor RIR, confirming that averaging raw RIRs does not preserve their acoustic structure. Performing the same averaging in the VAE latent space substantially reduces reconstruction error despite using identical geometric weighting. This result suggests that the learned latent representation is more amenable to interpolation than the time domain.

Among the kernel methods, increasing the RBF bandwidth shows a mild tendency toward improved performance, consistent with Table 2. In Fig. 1, $\text{RBF}_{0.2}$ generally performs better than narrower kernels as nearest-neighbor distance increases, while the opposite can occur at the smallest distances, most clearly in bin (0.1, 0.14] m of Fig. 1(b.iii). This behavior is consistent with the expected tradeoff between narrow kernels, which resemble nearest-neighbor interpolation, and wider kernels, which incorporate information from more reference RIRs.

Overall, these results suggest that training-free latent-space kernel interpolation is competitive across diverse acoustic conditions, with its advantage over nearest-neighbor estimation more apparent as nearest-reference distance increases. The preferred kernel bandwidth depends on both nearest-reference distance and the acoustic environment.

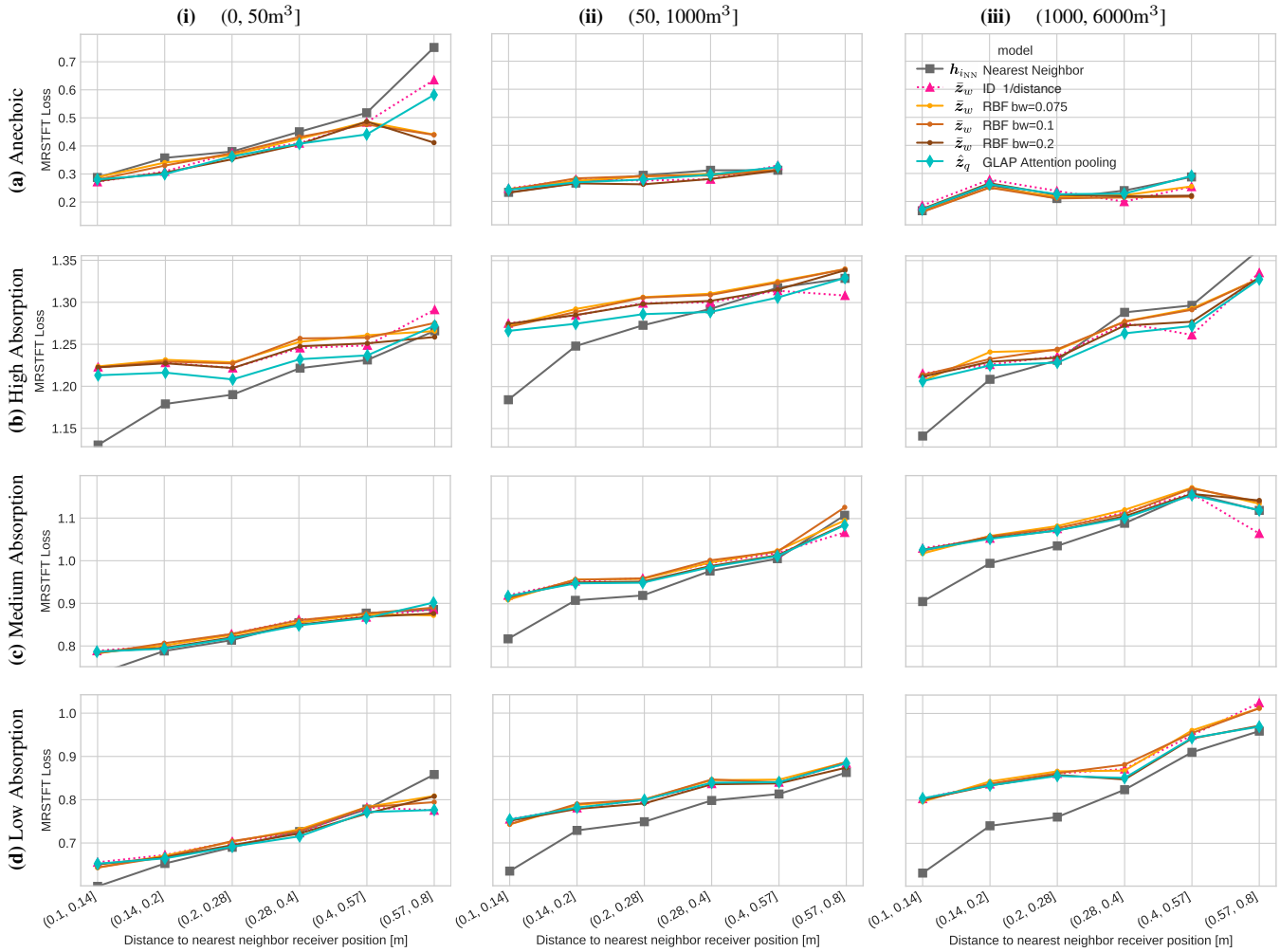

Fig. 1: MRSTFT error versus distance from the nearest reference position for the nearest neighbor, weighted average of latents (with inverse distance and RBF as kernel functions) and GLAP. Rows show material types from most to least absorbent, and columns show room volume.

5.2. Geometry-informed latent attention pooling results

The geometry-informed latent attention pooling model GLAP uses the same VAE latent domain, but replaces fixed distance-based weighting with learned query-conditioned aggregation. Compared with nearest neighbor and NW estimators with fixed kernels, the attention pooling model can adaptively select useful reference RIRs and apply a learned residual correction.

In the easier subsets, such as small or absorbent rooms in Fig. 1 (a.i), (b.i), and (b.ii), GLAP consistently improves upon the kernel methods. In larger and more reverberant rooms, however, this advantage diminishes, and GLAP becomes comparable to the strongest kernel baseline. This behavior is also reflected in Table 2, where GLAP achieves the lowest average latent reconstruction error and modest improvements in several decoded acoustic metrics.

These results indicate that the learned latent representation already provides a strong foundation for interpolation, with simple kernel weighting remaining competitive across most acoustic conditions. Within this representation, GLAP extracts modest additional improvements through decoder-aware training and query-dependent aggregation, particularly in smaller or less reverberant environments. The relatively small gains over the strongest kernel baselines suggest that the learned representation itself, rather than the interpolation strategy alone, is likely the dominant limitation.

6. CONCLUSION

In this work, we investigate few-shot RIR interpolation in a VAE latent domain. We demonstrate that simple weighted averaging in the latent domain yields reasonable interpolation performance, particularly as the distance to the nearest reference RIR increases. We also explored a lightweight re-bottlenecking model GLAP, which provides modest improvements over fixed kernel weighting under favorable acoustic conditions. The results highlight the potential of latent representations for few-shot RIR interpolation across diverse environments, and suggest that further gains depend on latent spaces with better reconstruction fidelity and physical consistency. Future work will investigate interpolation beyond convex-hull settings with more realistic room geometries, as well as extensions to binaural and spatially dynamic rendering scenarios.

7. REFERENCES

[1] S. Koyama, E. De Sena, P. Samarasinghe, M. R. P. Thomas, and F. Antonacci, "Past, present, and future of spatial audio and room acoustics," in *Proc. ICASSP*, 2025.

[2] M. Vorländer, D. Schröder, S. Pelzer, and F. Wefers, "Virtual reality for architectural acoustics," *J. of Build. Perform. Simul.*, vol. 8, no. 1, pp. 15–25, 2015.

[3] L. Bahrman, M. Fontaine, J. Le Roux, and G. Richard, "Speech dereverberation constrained on room impulse response characteristics," in *Proc. Interspeech*, 2024, pp. 622–626.

[4] R. Mignot, L. Daudet, and F. Ollivier, "Room reverberation reconstruction: Interpolation of the early part using compressed sensing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 11, pp. 2301–2312, 2013.

[5] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. Van Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1929–1941, 2017.

[6] N. Ueno, S. Koyama, and H. Saruwatari, "Kernel ridge regression with constraint of Helmholtz equation for sound field interpolation," in *Proc. IWAENC*, 2018.

[7] O. Das, P. Calamia, and S. V. A. Gari, "Room impulse response interpolation from a sparse set of measurements using a modal architecture," in *Proc. ICASSP*, 2021, pp. 960–964.

[8] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.

[9] A. Geldert, N. Meyer-Kahlen, and S. J. Schlecht, "Interpolation of spatial room impulse responses using partial optimal transport," in *Proc. ICASSP*, 2023, pp. 1–5.

[10] A. Geldert, "Room impulse response interpolation via optimal transport," Master's thesis, Aalto University, 2023.

[11] D. Sundström, F. Elvander, and A. Jakobsson, "Optimal transport based impulse response interpolation in the presence of calibration errors," *IEEE Trans. Signal Process.*, vol. 72, pp. 1548–1559, 2024.

[12] J. G. C. Ribeiro, S. Koyama, R. Horiuchi, and H. Saruwatari, "Sound field estimation based on physics-constrained kernel interpolation adapted to environment," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 4369–4383, 2024.

[13] A. Richard, P. Dodds, and V. K. Ithapu, "Deep impulse responses: Estimating and parameterizing filters with deep networks," in *Proc. ICASSP*, 2022, pp. 3209–3213.

[14] Y. Qiao and E. Choueiri, "Neural modeling and interpolation of binaural room impulse responses with head tracking," in *Audio Engineering Society Convention*, 2023.

[15] M. Eineborg, F. Pind, and S. Thrastarson, "Generating impulse responses using autoencoders," in *Proc. Forum Acusticum*, Jan. 2024, pp. 3177–3180.

[16] X. Karakonstantis, D. Caviedes-Nozal, A. Richard, and E. Fernandez-Grande, "Room impulse response reconstruction with physics-informed deep learning," *J. Acoust. Soc. Am.*, vol. 155, no. 2, pp. 1048–1059, 2024.

[17] Z. Lan, C. Zheng, Z. Zheng, and M. Zhao, "Acoustic volume rendering for neural impulse response fields," in *Proc. NeurIPS*, 2024.

[18] M. L. Wang, R. Sawata, S. Clarke, R. Gao, S. Wu, and J. Wu, "Hearing anything anywhere," in *Proc. CVPR*, 2024.

[19] S. Della Torre, M. Pezzoli, F. Antonacci, and S. Gannot, "DiffusionRIR: Room impulse response interpolation using diffusion models," in *Proc. Forum Acusticum / EuroNoise*, Dec. 2025, pp. 4079–4086.

[20] S. Koyama and K. Ishizuka, "Learning magnitude distribution of sound fields via conditioned autoencoder," *arXiv preprint arXiv:2506.16729*, 2025.

[21] A. Luo, Y. Du, M. J. Tarr, J. B. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," in *Proc. NeurIPS*, 2022.

[22] K. Su, M. Chen, and E. Shlizerman, "INRAS: Implicit neural representation for audio scenes," in *Proc. NeurIPS*, 2022.

[23] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu, "AV-NeRF: Learning neural fields for real-world audio-visual scene synthesis," in *Proc. NeurIPS*, 2023.

[24] S. Majumder, C. Chen, Z. Al-Halah, and K. Grauman, "Few-shot audio-visual learning of environment acoustics," in *Proc. NeurIPS*, 2022.

[25] X. Liu, A. Kumar, P. Calamia, S. V. Amengual, C. Murdock, I. Ananthabhotla, P. Robinson, E. Shlizerman, V. K. Ithapu, and R. Gao, "Hearing anywhere in any environment," in *Proc. CVPR*, 2025.

[26] E. A. Nadaraya, "On estimating regression," *Theory of Probability & Its Applications*, vol. 9, no. 1, pp. 141–142, 1964.

[27] G. S. Watson, "Smooth regression analysis," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372, 1964.

[28] D. Bralios, J. Casebeer, and P. Smaragdis, "Re-Bottleneck: Latent re-structuring for neural audio autoencoders," in *Proc. MLSP*, 2025.

[29] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Stable Audio Open," in *Proc. ICASSP*, 2025.

[30] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Proc. NeurIPS*, Dec. 2020, pp. 7537–7547.

[31] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulations and array processing algorithms," in *Proc. ICASSP*, Apr. 2018.