NEURAL ACOUSTIC MULTIPOLE SPLATTING FOR ROOM IMPULSE RESPONSE SYNTHESIS

*Geonwoo Baek[1] and Jung-Woo Choi[1]**

[1]School of Electrical Engineering, KAIST, Daejeon, Republic of Korea
{bkw6287, jwoo}@kaist.ac.kr

ABSTRACT

Room Impulse Response (RIR) prediction at arbitrary receiver positions is essential for practical applications such as spatial audio rendering. We propose Neural Acoustic Multipole Splatting (NAMS), which synthesizes RIRs at unseen receiver positions by learning the positions of neural acoustic multipoles and predicting their emitted signals and directivities using a neural network. Representing sound fields through a combination of multipoles offers sufficient flexibility to express complex acoustic scenes while adhering to physical constraints such as the Helmholtz equation. We also introduce a pruning strategy that starts from a dense splatting of neural acoustic multipoles and progressively eliminates redundant ones during training. Experiments conducted on both real and synthetic datasets indicate that the proposed method surpasses previous approaches on most metrics while maintaining rapid inference. Ablation studies reveal that multipole splatting with pruning achieves better performance than the monopole model with just 20% of the poles.

Index Terms— Room impulse response synthesis, neural acoustic multipole, pruning strategy

1. INTRODUCTION

Room Impulse Responses (RIRs) characterize how sound propagates in a room, including direct sound, early reflections, and late reverberation. Accurate RIR rendering is crucial for spatial audio, virtual and augmented reality, and interactive gaming, as it provides realistic and immersive room cues at various listener positions. In reality, we can only obtain a limited number of RIRs, but we need to simulate them for numerous unseen receiver locations. This paper tackles the fundamental challenge of predicting RIRs at unseen positions using a limited dataset of measured or simulated RIRs.

RIR estimation has been addressed through multiple approaches. Notable examples include modal expansion [1, 2], the equivalent source method (ESM) [3, 4, 5], and plane wave expansion (PWE) [5, 6, 7]-based methods, all aimed at solving the acoustic inverse problem using regularizations and constraints. These traditional techniques explicitly model physical structures to recreate sound fields. Yet, accuracy typically suffers at high frequencies, and practical application to wideband audio is hampered by ill-conditioning and aliasing artifacts. Additionally, RIR parameterization techniques [8, 9] have been introduced to model RIRs using several room acoustic parameters, which enable dynamic acoustic effects in gaming. However, these estimations can be suboptimal in addressing all aspects of RIRs.

Recently, deep neural networks (DNNs) have gained traction for this task. Initially, DNNs estimated RIRs from given source

and receiver positions [10, 11] but lacked physical consistency and struggled with high-frequency components. Other approaches introduced physical constraints to enhance model stability and generalization [12, 13, 14]. They integrated physical priors, like the wave equation, into the loss function to guide learning [12]. For better early reflection predictions, some models used periodic activation functions [13, 15], while others applied the dynamic pulling method to enhance performance in noisy settings [14]. These models need only a few RIRs for training [12] and ensure physical fidelity, yet training can be slow due to explicit PDE residual computation and may face gradient issues, affecting high-frequency accuracy [16].

Recently, approaches that incorporate pre-designed physical models into the network optimization have been suggested [17, 16]. These approaches develop a physical model inherently satisfying necessary physical constraints and estimate model inputs or parameters that align with the training data. Such approaches offer improvements over previous methods, particularly in capturing high-frequency components [16]. For example, AVR [17] models the sound field as the combination of signals emitted from points on rays surrounding the receiver and considers the attenuation and time delay caused by sound propagation through rays. The point neuron model [16] uses point neurons satisfying the Helmholtz equation as basis functions to describe the sound field and trains their signals and positions to encode and estimate the entire field. These model-fitting approaches aligning the basis functions to the observed signal are in line with Gaussian Splatting (GS) [18] that utilizes Gaussian functions for 3-D field reconstruction from 2-D vision images.

Existing acoustic model-fitting methods often suffer from inefficiencies of basis functions like points on rays or acoustic monopoles, which require a large number to function effectively, thereby increasing inference time. For instance, the image source method [19] can effectively mimic early reflections from specular surfaces using monopoles but struggles with simulating scattering from walls and objects. Such irregular reflections can be better represented as directional multipoles. While compact monopoles can also act as directional multipoles, achieving high-order directivities leads to amplified monopole signals, causing instability in DNN optimizations.

To address this, we introduce a multipole-based RIR estimation network, NAMS, which uses acoustic multipoles with adjustable directivities to model and estimate RIRs. This model spatially distributes multipoles, optimizes their positions, and predicts both their signals and frequency-dependent directivities. In addition, inspired by adaptive density control of GS [18], we propose a pruning strategy that optimizes the number of multipoles during training. Initially, we densely splat multipoles and progressively prune redundant ones throughout training. Our experimental results conducted in real and virtual environments reveal that NAMS outperforms existing methods in terms of estimated room acoustic parameters while achieving faster inference speeds.
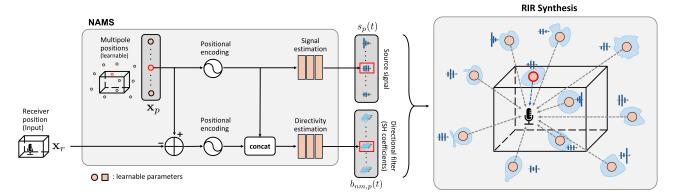
---
*Corresponding author.

Fig. 1: **Overview of NAMS framework.** We optimize each neural acoustic multipole position $\mathbf{x}_p$ and train MLPs to predict its emitted signal $s_p(t)$ and spherical harmonic coefficients $b_{nm,p}(t)$. We synthesize the RIR using the receiver position $\mathbf{x}_r$ and the set of neural acoustic multipole positions, emitted signals, and spherical harmonic coefficients.

2. PROPOSED METHOD

2.1. NAMS for Room Impulse Response Synthesis

NAMS generates RIRs by employing multipoles that have adjustable directional patterns and emit signals of limited duration. The directional patterns $D_p(f)$ for the p-th multipole at frequency f can be expressed in terms of the spherical harmonics [20] as

$$D_p(f, \mathbf{x}_r) \;=\; \sum_{n=0}^{N} \sum_{m=-n}^{n} B_{nm,p}(f)\, Y_n^m(\boldsymbol{\Omega}_p(\mathbf{x}_r)), \quad (1)$$

where $n \leq N$ and m denote the order and degree of spherical harmonics Y_n^m, and $\boldsymbol{\Omega}_p(\mathbf{x}_r)$ indicates the angular position of the multipole at $\mathbf{x}_p$ measured from the receiver position $\mathbf{x}_r$. The coefficients $B_{nm,p}$ determine the directional pattern of the p-th multipole.

We model the RIR as the superposition of these multipoles, each of which emits a signal $S_p(f)$. With a far-field assumption, an RIR in frequency and time domains can be approximated as

$$H(f, \mathbf{x_r}) \;=\; \sum_{p=1}^{P} S_p(f) \frac{e^{-j2\pi f r_p(\mathbf{x_r})/c}}{r_p(\mathbf{x_r})}\, D_p(f, \mathbf{x_r}), \quad (2)$$

where $e^{-j2\pi f r_p/c}$ expresses the propagation delay for the speed of sound c, and $1/r_p$ represents attenuation across the propagation distance $r_p(\mathbf{x}_r) = \|\mathbf{x}_r - \mathbf{x}_p\|$.

Our model predicts each multipole position $\mathbf{x}_p$, its emitting signal $s_p(t) = \mathcal{F}^{-1}[S_p(f)]$, as well as its spherical harmonic coefficients $b_{nm,p}(t) = \mathcal{F}^{-1}[B_{mn,p}(f)]$, for the given receiver position $\mathbf{x}_r$ and inverse Fourier transform $\mathcal{F}^{-1}$. The RIR is then synthesized in the frequency domain from Eq. (2) using a set of estimated parameters $\Theta = \{(\mathbf{x}_p, S_p(f), B_{nm,p}(f))\}$. The model is trained to minimize a loss function designed to compare the ground truth and estimated RIRs (detailed loss functions are described in Section 3.2).

The NAMS architecture to estimate the parameter Θ is depicted in Fig. 1. The architecture consists of signal and directivity branches for estimating multipole position and signal pairs $\{\mathbf{x}_p, s_p(t)\}_{p=1}^{P}$ and directional patterns $\{D_p(f)\}$, respectively. In the upper branch, a set of multipole positions $\{\mathbf{x}_p\}$ is declared as learnable network parameters, and only this information is utilized to synthesize the multipole signal $s_p(t)$. This process is to ensure that the multipole signals are independent of receiver positions. Specifically, the multipole position is encoded by a positional encoder and then fed into

the signal estimation layer (MLP) to generate a short-length signal $s_p(t)$ in the time domain. The multipole index p is assigned to the batch dimension, and MLP is trained to generate source signals corresponding to their multipole positions. In the lower branch, the model takes the receiver position $\mathbf{x}_r$ as input and subtracts it from multipole positions $\mathbf{x}_p$ to calculate the relative positions of multipoles required for deriving $\Omega_p(\mathbf{x}_r)$. The relative multipole positions are also encoded by a positional encoder and concatenated with the encoded $\mathbf{x}_p$. From the encoded position vectors, the directivity estimation layer (MLP) synthesizes the spherical harmonic coefficients $\{b_{nm,p}(t) = \mathcal{F}^{-1}[B_{nm,p}(f)]\}$ in the time domain.

The coefficients are then used to generate the directional patterns according to Eq. (1). We use real-valued spherical harmonics and constrain the total energy of $D_p(f, \mathbf{x}_r)$ across frequencies to stabilize the training process. In detail, for a vector $\mathbf{D}_p = [D_p(f_1, \mathbf{x}_r), \cdots, D_p(f_F, \mathbf{x}_r)]$ defined for discrete frequencies $f_1, \cdots, f_F$, we apply normalization $\hat{\mathbf{D}}_p = \mathbf{D}_p/\|\mathbf{D}_p\|$ and use it in place of $D_p(f, \mathbf{x}_r)$ of Eq. (2). The entire process described above is differentiable, so we can train both $\{\mathbf{x}_p\}_{p=1}^{P}$ and the MLPs via backpropagation. Moreover, the physical constraint on the Helmholtz equation is automatically satisfied, because multipoles constituting Eq. (2) are the solution of the Helmholtz equation.

2.2. Pruning Neural Multipoles

An excessive number of multipoles can lead to overfitting, significantly degrading computational efficiency. Therefore, it is essential to determine the optimal number of multipoles. To achieve this, we densely distribute the multipoles in the space at model initialization and incorporate pruning stages during training. We regularly perform pruning every 20 epochs after the first 100 epochs of training. The principle of pruning is to remove multipoles with low $s_p(t)$ energy. To compute the energy of $s_p(t)$ for each multipole, we freeze the model and input a dummy receiver position. Since $s_p(t)$ depends solely on $\mathbf{x}_p$, the choice of dummy receiver position does not affect the result. Also, the total energy of $D_p(f)$ is constrained to one, so we can determine the necessity of each multipole solely by the energy of $s_p(t)$. If the calculated energy is below 50% of the global median, the corresponding $\mathbf{x}_p$ is removed from the model. After each pruning step, the model re-optimizes $\{\mathbf{x}_p\}_{p=1}^{P'}$ and estimation layers using the reduced number (P') of multipoles. This iterative alternating process allows for the development of a compact NAMS model with less possibility of overfitting.

3. EXPERIMENT

3.1. Datasets

Our model's performance was assessed using both real and synthetic datasets. For the real dataset, we utilized the MeshRIR dataset [21]. As for the synthetic dataset, we conducted simulations of two room environments employing the Treble simulator [1], which offers a hybrid simulation approach integrating wave-based and geometrical acoustics. Specifically, we chose two scenes: Apartment 566 and Apartment 716 from the Treble database. The scenes contain furniture such as sofas, carpets, and beds, with absorption and scattering coefficients taken from the predefined values in the Treble database. Apartments 566 and 716 have volumes of 105 m^3 and 183 m^3, respectively, and follow a Manhattan layout. The reverberation times (T60) of Apartments 566 and 716, averaged for all receiver positions, were 0.48 s and 1.80 s, respectively. We positioned a single omnidirectional source at a fixed location selected at random. RIRs were simulated for 1,000 randomly chosen receiver positions. Fig. 2 depicts the employed room layouts. Across all experiments, we randomly split the RIRs into training and testing sets with a ratio of 9:1. All RIRs were resampled to the 24 kHz sampling rate and trimmed at 0.1 seconds.
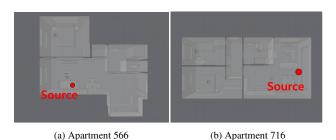

(a) Apartment 566   (b) Apartment 716

Fig. 2: Room and source configurations used for the simulations.

3.2. Implementation Details

During the initial phase, we densely splatted multipoles in space. The starting locations were arranged as points on a collection of spheres, each centered on the source, with radii incrementing from 1 m to 34 m at intervals of 1 m. On each sphere, 32 points were uniformly distributed using Fibonacci sampling and randomly rotated together. An additional multipole was located at the source position to render the direct sound, yielding a total of 1,089 initial positions. We applied sinusoidal positional embedding with 10 sine–cosine frequencies to encode source and receiver positions, following [10]. A 3-layer MLP with 512 hidden units per layer was used to predict $s_p(t)$, and the same architecture was employed to predict $b_{nm,p}(t)$. The output $s_p(t)$ has a duration of 3 ms, and $b_{nm,p}(t)$ includes spherical harmonic coefficients up to the 3rd order, yielding a 16-channel output with a duration of 3 ms.

For optimization, we adopted the loss function used in AVR [17], which consists of a weighted sum of spectral loss, amplitude loss, phase loss, time-domain loss, multi-resolution STFT loss [22], and energy decay loss [23], with weights of 1, 0.5, 0.5, 100, 1, and 5, respectively. We used the Adam optimizer with a cosine scheduler, decaying the learning rate from 10^{-3} to 10^{-4}. All experiments were conducted for 300 epochs, and the best model was updated at the epoch whenever the test loss was minimized. All experiments were run on a single RTX A6000 GPU.
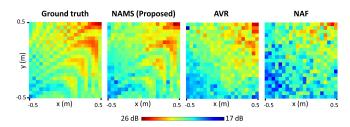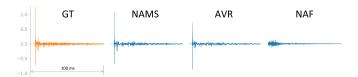
---

[1] https://www.treble.tech

Fig. 3: **Comparison of spatial magnitude distribution**. Overall magnitude distribution averaged over the 1/3 octave band centered at 4 KHz (Scene from MeshRIR).

Fig. 4: **Ground-truth vs. generated RIRs.** Samples with comparable amplitude errors to Table 1 are selected from MeshRIR.

4. RESULTS

4.1. Performance Comparison

We compared our model with NAF [10] and AVR [17] that estimate RIRs only from the given source and receiver positions. For a fair comparison, all models were trained for the same source position in the real and synthetic datasets. The point-neuron model [16] was excluded from comparison because its code is not publicly available. Instead, in our ablation study, we evaluated performance differences between monopole and multipole configurations.

The model performances were assessed using the following metrics: phase error, amplitude error, envelope error, reverberation time (T60) error, clarity (C50) error, and early decay time (EDT) error [17]. Phase and amplitude errors are relative errors with no unit; envelope error is presented as a relative percentage error, and the units for T60, C50, and EDT errors are %, dB, and milliseconds, respectively.

We present the experimental results in Table 1. On the MeshRIR dataset, NAMS outperforms existing methods on all metrics. Similar trends can be seen on the Apartment 566 and 716 datasets, except for the envelope error. However, all models yield a phase error of approximately 1.62 on the synthetic dataset, probably because the rapid phase increase from large propagation delay is wrapped within $\pm\pi$. To visualize the differences in the estimated sound field, we present the spatial magnitude distribution of sound fields estimated from the MeshRIR dataset. The results show a clear difference of NAMS compared to baseline models with less spatial jitter. In terms of RIR waveforms shown in Fig. 4, NAMS captures the direct sound and early reflection peaks more accurately than AVR and NAF.

In Table 3, we also compare the space and time complexity of models. NAMS also shows a fast inference time (T_{Inf}) of 2.1–2.2 ms, comparable to NAF (1.9–2.0 ms). NAMS achieves this speed using only a few hundred multipoles, whereas AVR requires sampling 204,800 points. In addition, our model has an advantage in terms of parameter size. AVR [17] has 57.2 million parameters, whereas NAMS and NAF [10] have 1.8 million and 2.7 million parameters, respectively. However, the large parameter count of AVR [17] is mainly due to its use of hash grid encoding [24]. Excluding the hash

Method	MeshRIR						Apartment 566						Apartment 716					
	Phase	Amp.	Env.	T60	C50	EDT	Phase	Amp.	Env.	T60	C50	EDT	Phase	Amp.	Env.	T60	C50	EDT
NAF[10]	1.62	0.57	1.98	3.5	0.88	31.0	1.62	0.77	5.05	15.2	8.66	22.0	1.62	0.76	6.56	21.4	7.35	18.8
AVR[17]	1.05	0.28	1.44	2.9	0.66	19.4	1.62	0.46	**4.20**	5.0	1.20	29.8	1.62	0.54	**5.79**	9.0	2.46	24.2
NAMS	**0.80**	**0.11**	**1.21**	**2.0**	**0.34**	**9.8**	**1.60**	**0.22**	4.46	**3.4**	**0.48**	**12.0**	**1.60**	**0.30**	6.70	**6.3**	**0.82**	**13.7**

Table 1: **RIR estimation performance comparison with existing models**. The evaluation metrics include the phase error (Phase.), amplitude error (Amp.), envelope error (Env., %), relative T60 error (%), C50 error (dB), and EDT error (ms).

Method	MeshRIR								Apartment 566								Apartment 716							
	Phase	Amp.	Env.	T60	C50	EDT	#pts	T_{Inf}	Phase	Amp.	Env.	T60	C50	EDT	#pts	T_{Inf}	Phase	Amp.	Env.	T60	C50	EDT	#pts	T_{Inf}
mono. sparse	0.93	0.19	1.36	2.3	0.39	12.4	273	**1.7**	1.61	0.26	4.81	4.0	0.60	16.0	307	**1.7**	1.62	0.59	10.25	7.8	0.82	16.1	273	**1.7**
multi. sparse	0.84	0.15	1.30	2.1	0.37	11.5	273	2.1	**1.60**	**0.22**	4.50	3.6	**0.43**	**11.5**	307	2.2	1.61	0.40	8.20	**5.9**	**0.76**	**13.5**	273	2.1
mono. dense	0.83	0.15	1.27	2.2	0.37	10.8	1089	2.0	1.61	0.28	4.89	4.0	0.69	17.0	1089	2.0	1.62	0.50	9.33	7.2	0.99	16.4	1089	2.0
multi. dense	**0.78**	**0.12**	**1.20**	**2.0**	**0.33**	10.2	1089	4.5	**1.60**	**0.22**	**4.39**	3.9	0.50	13.3	1089	4.6	1.61	0.42	8.47	6.3	0.87	13.6	1089	4.6
multi. dense *w/ pruning*	0.80	**0.11**	1.21	**2.0**	0.34	**9.8**	**225**	2.2	**1.60**	**0.22**	4.46	**3.4**	0.48	12.0	**276**	2.2	**1.60**	**0.30**	**6.70**	6.3	0.82	13.7	**240**	2.1

Table 2: **Efficacy of multipole and pruning.** Performance comparison of monopole and multipole splatting. 'mono.' and 'multi.' represent monopole and multipole settings, respectively, whereas 'sparse' and 'dense' denote the small and large number of poles at initialization.

Method	MeshRIR			Apartment 566 & 716		
	Param.	#pts	T_{Inf}	Param.	#pts	T_{Inf}
NAF[10]	2.7M	-	**1.9**	2.7M	-	**1.95**
AVR[17]	57.2M (2.0M)	205k	62.5	57.2M (2.0M)	205k	61.9
NAMS	**1.8M**	**225**	2.2	**1.8M**	**258**	2.15

Table 3: **Space and time complexity** Comparison of the model's parameter size (Param.), number of sample points or multipoles (#pts), and inference time (T_{Inf}, ms).
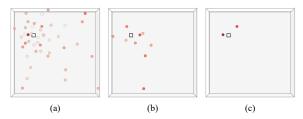

Fig. 5: **Multipole positions and signal energies with and without pruning.** Circles indicate the multipoles inside the room, with deeper red indicating higher energy. The square denotes the source. (a) dense w/o pruning, (b) sparse w/o pruning, (c) dense w/ pruning.

grid encoding, AVR has a slightly higher number of parameters.

4.2. Ablation studies and discussions

We conducted ablation studies to investigate the efficacy of multipole splatting, by comparing the monopole and multipole models. The monopole model is implemented by setting the maximum order (N) of spherical harmonics to zero. We also considered two different pole initializations for each configuration. The dense initialization denotes the initialization described in Section 3.2, while the sparse initialization indicates the case where the number of points on each sphere is reduced such that the total number of poles is similar to that obtained after pruning the dense initialization model. The results are shown in Table 2.

For the same number of poles, multipole models consistently outperform monopole models across all datasets. Despite this, mul-tipole models incur higher inference time due to their larger number of trainable parameters. Nevertheless, in both the Apartment 566 and 716 datasets, the sparse multipole model outperforms all monopole models, including the dense initialization model. This result indicates that multipoles offer a more expressive representation in complex acoustic environments. Conversely, in the MeshRIR dataset, both the multipole model with sparse initialization and the monopole model with dense initialization perform similarly, suggesting that monopoles might suffice for modeling uncomplicated, unobstructed rooms.

As indicated in Table 2, pruning decreased the multipoles to 225, 276, and 240 in the MeshRIR, Apartment 566, and Apartment 716 datasets, respectively, yet still outperformed the dense monopole model with 1089 poles. This highlights that the multipole model more effectively represents sound fields using only 20–22% of poles. The pruned model also resulted in over twice the inference speed compared to the non-pruned dense multipole initialization for each dataset. Moreover, the compact model aids in interpreting the physical structure of a sound field. To illustrate, we present the multipole distributions optimized for the MeshRIR dataset under three scenarios (Fig. 5): dense and sparse splatting without pruning, and dense splatting with pruning. In the pruned dense setup, two high-energy multipoles are placed near the source, whereas both dense and sparse setups without pruning spread many low-energy multipoles in the space. Thus, the pruning model offers a more interpretable, structured portrayal of a sound field using a limited set of multipoles with adaptable directivities and signal emission.

5. CONCLUSION

We introduced NAMS to synthesize room impulse responses via multipoles, using pruning to auto-select the optimal quantity. This approach outperforms existing methods while ensuring quick inference by efficiently representing sound fields with refined multipole directivities and placements. We demonstrated that multipoles offer a richer representation than monopoles in complex acoustic scenarios. Additionally, starting with a dense multipole set and optimizing through pruning achieves better results than manual initialization. This evidence shows NAMS efficiently represents RIRs. The next step for practical RIR estimation will be designing a generalized model for various source positions using fewer RIRs.

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] Y. Haneda, Y. Kaneda, and N. Kitawaki, "Common-acoustical-pole and residue model and its application to spatial interpolation and extrapolation of a room transfer function," *IEEE/ACM Transactions on Speech and Audio Processing (TASLP)*, vol. 7, no. 6, pp. 709–717, 1999.

[2] O. Das, P. Calamia, and S. Gari, "Room impulse response interpolation from a sparse set of measurements using a modal architecture," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 960–964.

[3] I. Tsunokuni, K. Kurokawa, H. Matsuhashi, Y. Ikeda, and N. Osaka, "Spatial extrapolation of early room impulse responses in local area using sparse equivalent sources and image source method," *Applied Acoustics*, vol. 179, pp. 108027, 2021.

[4] H. Matsuhashi, I. Tsunokuni, and Y. Ikeda, "Spatial Interpolation of Early Room Impulse Responses Using Equivalent Source method based on Grouped Image Sources," in *Proc. INTER-NOISE and NOISE-CON Congress and Conference*. Institute of Noise Control Engineering, 2023, vol. 265, pp. 4891–4897.

[5] Niccolo Antonello, Enzo De Sena, Marc Moonen, Patrick A Naylor, and Toon Van Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1929–1941, 2017.

[6] M. Hahmann and E. Fernandez-Grande, "A convolutional plane wave model for sound field reconstruction," *The Journal of the Acoustical Society of America (JASA)*, vol. 152, no. 5, pp. 3059–3068, 2022.

[7] W. Jin and W. Kleijn, "Theory and design of multizone sound-field reproduction using sparse methods," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2343–2355, 2015.

[8] N. Raghuvanshi and J. Snyder, "Parametric directional coding for precomputed sound propagation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[9] N. Raghuvanshi and J. Snyder, "Parametric wave field coding for precomputed sound propagation," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 1–11, 2014.

[10] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022, vol. 35, pp. 3165–3177.

[11] K. Su, M. Chen, and E. Shlizerman, "Inras: Implicit neural representation for audio scenes," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022, vol. 35, pp. 8144–8158.

[12] X. Karakonstantis, D. Caviedes-Nozal, A. Richard, and E. Fernandez-Grande, "Room impulse response reconstruction with physics-informed deep learning," *The Journal of the Acoustical Society of America (JASA)*, vol. 155, no. 2, pp. 1048–1059, 2024.

[13] M. Pezzoli, F. Antonacci, and A. Sarti, "Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses," *arXiv preprint arXiv:2306.11509*, 2023, [Online]. Available: https://arxiv.org/pdf/2306.11509.

[14] K. Kurata, G. Sato, I. Tsunokuni, and Y. Ikeda, "Noise-Robust Estimation of Early-part Room Impulse Responses based on Physics-Informed Neural Network with Dynamic Pulling Method," in *Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2024, pp. 1–5.

[15] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020, vol. 33, pp. 7462–7473.

[16] H. Bi and T. Abhayapala, "Point neuron learning: a new physics-informed neural network architecture," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, pp. 56, 2024.

[17] Z. Lan, C. Zheng, Z. Zheng, and M. Zhao, "Acoustic volume rendering for neural impulse response fields," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024, vol. 37, pp. 44600–44623.

[18] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering.," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 139–1, 2023.

[19] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America (JASA)*, vol. 65, no. 4, pp. 943–950, 1979.

[20] N. Gumerov and R. Duraiswami, *Fast multipole methods for the Helmholtz equation in three dimensions*, Elsevier, 2005.

[21] S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, and J. Brunnström, "MeshRIR: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods," in *Proc. IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*. IEEE, 2021, pp. 1–5.

[22] R. Yamamoto, E. Song, and J. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[23] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Few-shot audio-visual learning of environment acoustics," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022, vol. 35, pp. 2522–2536.

[24] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.