TA-RIR: Topology-Aware Neural Modeling of Acoustic Propagation for Room Impulse Response Synthesis

Junhui Zhao[1], *Hang Chen*[1], *Qing Wang*[1], *Jun Du*[1], *Yanhui Tu*[2], *Feng Ma*[2]

[1]Department of Electronic Engineering and Information Science, University of Science and Technology of China, China
[2]RDG Group, iFlytek Research, China

jhzhao@mail.ustc.edu.cn, hangchen@ustc.edu.cn

Abstract

Accurate estimation of room impulse responses (RIRs) is crucial for applications like augmented reality and sound field modeling. Current methods either neglect the spatial relationships between the source and receiver or rely on computationally intensive volumetric grids or panoramic images to estimate RIRs. To address these challenges, we introduce TA-RIR, a topology-aware neural network that uses spatial coordinates of sources and receivers, along with reverberant speech, to learn compact embeddings encoding room geometry and acoustics. The topology-aware encoder captures structural relationships between spatial and acoustic features, integrated through a propagation-informed decoder to synthesize RIRs. Experimental results show that TA-RIR generates high-fidelity RIRs, accurately preserving target acoustic parameters such as reverberation time, while significantly reducing computational complexity compared to methods requiring detailed 3D models or room acoustic properties.

Index Terms: room impulse response, synthesis, speech-simulation, topological embedding

1. Introduction

The room impulse response plays a crucial role in various research areas, such as far-field speech recognition [1], speech dereverberation [2], and augmented reality (AR) [3]. It provides a detailed representation of the acoustic relationship between a sound source and a microphone within a specific room. Several factors influence the RIR, including the room's geometry, the materials present, the sound source, and the position of the receiver. RIR can be measured by recording the excitation signal emitted by the sound source through the receiver, followed by deconvolution techniques. [4]. However, this direct measurement approach requires high-fidelity sound acquisition equipment and a controlled, noise-free environment, both of which are costly and limit the measurement to a specific sound source and receiver, thereby restricting its applicability in many downstream tasks. Consequently, researchers have explored estimating environmental acoustic parameters such as Reverberation Time (RT_{60}), early-to-late index (CTE), and Direct-to-Reverberant Ratio (DRR). These parameters offer insights into the environment's acoustic characteristics. For instance, Falcon et al. [5] employed a machine learning-based method to estimate room acoustic parameters from geometric information. However, for applications like AR, which require detailed acoustic information, estimating only these parameters may be insufficient, direct estimation of RIR is a preferable way.

Various methods have been developed to simulate RIRs directly, including wave-based approaches [6] and geometric methods [7, 8]. However, many of these techniques are computationally intensive. Recently, neural network-based methods have gained prominence. Pezzoli et al. [9] introduced a depth prior method to predict missing parts in RIRs measured by a uniform linear microphone array. Ratnarajah et al. proposed the TS-RIR [10], which compensates for low-frequency wave effects by using a Generative Adversarial Network (GAN) to convert synthetic RIRs into realistic ones. Martin et al. [11] developed an end-to-end encoder-decoder model to estimate RIRs from real RIRs, position information, and acoustic parameters.

Many existing methods for RIR estimation rely on reverberant speech signals [12, 13], but they often neglect the spatial relationship between the sound source and the receiver within the room. Other techniques incorporate additional modalities, such as visual information, room geometry, and material acoustic properties, to estimate RIRs. For instance, some approaches use indoor 3D scene models and the locations of the sound source and receiver to estimate the RIR. Ratnarajah et al. [14] utilized panoramic images of the room and the acoustic properties of materials to improve RIR estimation. While these methods can enhance RIR estimation accuracy, they often require additional information, such as 3D models and detailed acoustic characteristics of the room, which can be challenging to obtain.

In this paper, we propose a novel approach that leverages the spatial relationship between the sound source, receiver, and room topology, without relying on large-scale models or difficult-to-acquire data. Our end-to-end deep learning framework uses reverberant speech, along with sound source and receiver positions and room vertex information, to estimate the RIR. Specifically, we employ a Graph Neural Network (GNN) to model the spatial topology, a cascaded time-domain encoder to extract RIR features from reverberant speech, and a decoder to reconstruct the RIR.

The remainder of this paper is organized as follows. Section 2 introduces the proposed framework, comprising an encoder-decoder architecture that jointly processes reverberant speech signals and positional topology information, along with a multi-resolution STFT loss function. Section 3 details the experimental protocol, including dataset configuration, implementation specifications, and evaluation metrics, followed by a comprehensive presentation of comparative results and systematic analysis of performance variations. Section 4 concludes the paper.

2. Proposed Method

The architecture of the proposed model is depicted in Figure 1. Building upon the baseline encoder-decoder framework introduced in [12], our system incorporates a novel topological embedding generation module. The model addresses the acoustic simulation challenge of predicting room impulse eesponses

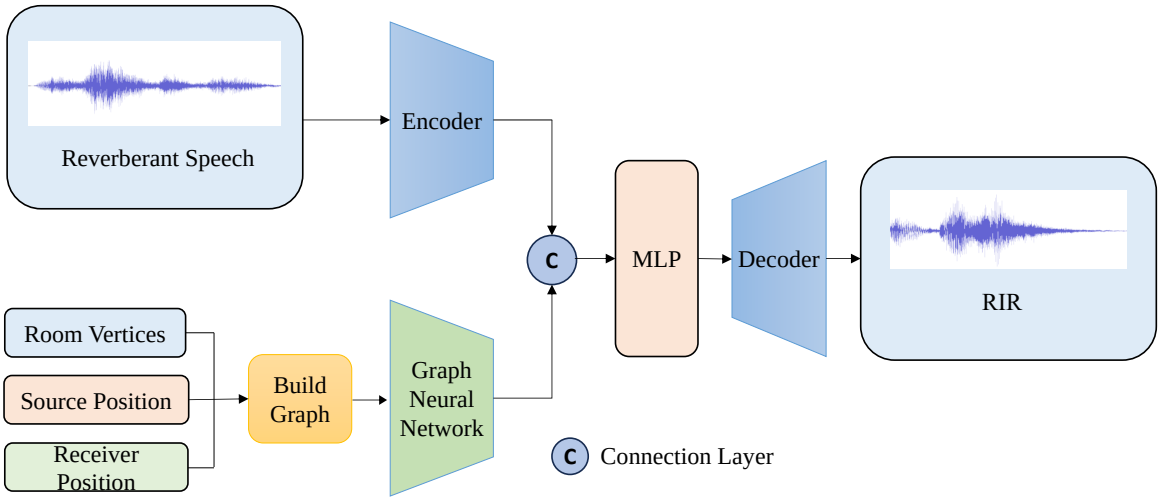

Figure 1: *The proposed method.*

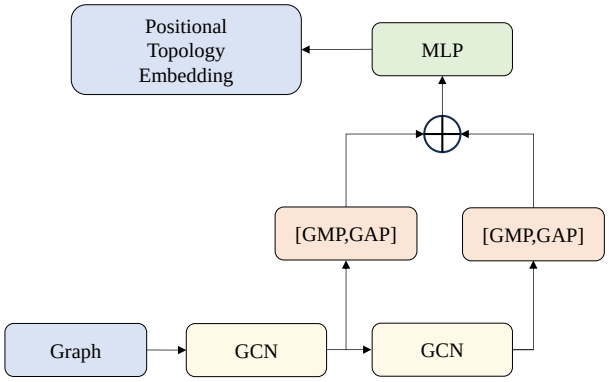

Figure 2: *Positional topology encoder structure.*

through a combination of spatial and acoustic feature learning, where inputs include three-dimensional coordinates of room vertices, sound source locations, and receiver positions. Specifically, the acoustic environment is geometrically represented as a polyhedral prism. The processing pipeline initiates with constructing a graph representation of the spatial topology, where vertices correspond to geometric coordinates and edges encode relative positional relationships. This graph structure undergoes feature extraction through our proposed Graph Neural Network (GNN) architecture to produce the topological embedding. Concurrently, the reverberant speech signal $x(l)$ undergoes acoustic feature transformation through the decoder structure adapted from [12]. The system subsequently employs a conditional generative architecture where the decoder synthesizes the RIR waveform from an initial noise vector. A fusion module is used to orchestrate the integration of topological and acoustic embeddings. This integration leverages Feature-wise Linear Modulation (FiLM) [15] to adaptively condition the RIR generation process, applying affine transformations to intermediate feature maps based on the combined embedding vector. The modulated features progressively refine the RIR prediction through successive upsampling operations in the decoder pathway.

2.1. Positional Topology Encoding Module

We utilize a graph neural network, as illustrated in Figure 2, to model the topological relationships between the positions of the sound source and the receiver within a room. We construct an undirected graph containing all the vertices of the room, the sound source and the receiver. The set of edges includes direct connections between the sound source and the receiver, along with edges linking the sound source and receiver to each individual vertex. Each vertex's feature is represented by its three-dimensional spatial coordinates. The proposed GNN architecture incorporates a graph convolution (GCN) layer [16] to encode the features of the graph structure data. This design aims to capture two dimensions of key information at the same time: first, the graph topology, which characterizes the relationships between nodes (particularly through the edge connections); and second, the node features, which represent the spatial coordinates of the vertices.

The layer-wise propagation rule of the graph convolutional layer is as follows [16]:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \quad (1)$$

Where $\tilde{A} = A + I$, A is the adjacency matrix of the graph. $\tilde{D}_{ij} = \sum_j \tilde{A}_{ij}$ represents the degree matrix. $W^{(l)}$ denotes the learnable weights of the l-th layer, and $H^{(l)}$ is the feature matrix of the lth layer. To extract the embedding vector for the decoder input, we apply both global average pooling (GAP) and global maximum pooling (GMP) to the feature maps obtained from each graph convolutional layer. These pooled features are then fused and passed through a linear layer. The dimension of the resulting position-topology embedding vector, denoted as π_P, is set to 64.

2.2. Reverberant Speech Encoder

The reverberant speech encoder is designed to effectively process time-domain reverberant speech signals. It comprises a series of residual encoder blocks. Each block initiates with a one-dimensional convolutional layer (kernel size = 15, stride = 2) with symmetric padding to analyze local acoustic patterns. Following this, batch normalization is applied, and the

Parametric Rectified Linear Unit (PReLU) activation function is employed [17] to introduce non-linearity and improve model expressiveness. To support deeper network architectures and mitigate the issue of vanishing gradients, each block incorporates a 1×1 one-dimensional convolutional layer followed by batch normalization, serving as a residual connection. A progressive channel expansion strategy is implemented, wherein the number of channels in each convolutional layer increases gradually as the network deepens. At the conclusion of the convolutional block sequence, features are aggregated using an adaptive average pooling layer and subsequently projected into a 128-dimensional time-domain reverb speech embedding, denoted as z_r, via a linear layer.

2.3. Fused Embedding Vector

The speech embedding z_r generated by the reverberant speech encoder is concatenated with the positional topology embedding z_p produced by the positional topology encoder. This concatenated vector is subsequently projected into a 64-dimensional space using a MLP. serving as the conditioning vector for the FiLM layer in the decoder.

2.4. Decoder

This model employs a decoder architecture that is isomorphic to the one described in the literature [12]. Each decoder module consists of two sequential processing stages. In the first stage, the fused embedding vector z is injected into the decoder network through the FiLM layer [15]. This layer maps the conditional embedding to modulation parameters γ (gain coefficient) and β (bias term) via an affine transformation, and applies a channel-by-channel affine transformation to the intermediate feature tensor along the channel dimension. This conditional modulation mechanism allows the model to dynamically incorporate global context information into the generation process, while preserving end-to-end differentiability during training, thereby offering refined control over the acoustic impulse response generation.

In terms of acoustic modeling, the impulse response produced by the model is explicitly decomposed into physically interpretable components: direct sound, early reflections $d(n)$, and late reverberation tail. The decoder directly generates the direct sound and early reflection components, along with their corresponding time-frequency mask matrices. These matrices are then applied to a set of predefined bandpass-filtered noise floor signals through element-wise multiplication to synthesize the late reverberation components. This hybrid generation strategy, enhances the acoustic realism of the generated results by employing a physically inspired signal decomposition method, while maintaining parameter efficiency.

2.5. Loss Function

We employ a multi-resolution short-time Fourier transform (STFT) loss [18] as the minimization criterion, which combines spectral convergence loss L_{MAG} and logarithmic STFT magnitude loss L_{SC}. This loss function enables joint optimization of the speech enhancement model at two levels: global spectral energy distribution and local time-frequency point details. The hierarchical feature matching enhances the signal reconstruction quality in the time-frequency domain. The definitions of these terms are as follows:

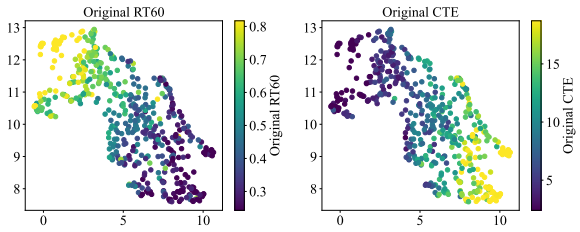

Figure 3: *2-D projections of fusion embeddings, colored by ground RT_{60} (left) and CTE (right).*

$$\mathcal{L}_{\mathrm{SC}}(\hat{x}, x) = \frac{\| |\mathrm{STFT}(x)| - |\mathrm{STFT}(\hat{x})| \|_F}{\| |\mathrm{STFT}(x)| \|_F} \tag{2}$$

$$\mathcal{L}_{\mathrm{MAG}}(\hat{x}, x) = \frac{1}{N} \| \log |\mathrm{STFT}(x)| - \log |\mathrm{STFT}(\hat{x})| \|_1, \tag{3}$$

$$\mathcal{L}_{\mathrm{STFT}}(\hat{x}, x) = \sum_{r=1}^{R} L_{\mathrm{SC}}(\hat{x}, x) + L_{\mathrm{MAG}}(\hat{x}, x) \tag{4}$$

Where x and $\hat{x}$ represent the original and generated IRs, respectively, and N denote the total number of STFT frames. The Frobenius norm and L_1 norm are denoted by $\|\cdot\|_F$ and $\|\cdot\|_1$, respectively. The spectral convergence term aims to constrain the overall energy distribution in the time-frequency representation by calculating the relative difference, in Frobenius norm, between the STFT amplitude spectrum of the predicted and target signals. The amplitude spectrum difference term uses the L_1 norm in the logarithmic scale to improve the model's ability to capture detailed spectral features. To create a multi-resolution representation, we calculate both the spectral convergence and amplitude spectrum difference terms at R different STFT resolutions, and combine them to form the final composite loss function. This multi-scale architecture effectively balances resolution in both the time and frequency domains, enhancing the reconstruction accuracy of the generated signal in the joint time-frequency domain.

3. Experiment and Results

3.1. Experiment Setup

To assess the effectiveness of the position topology encoding module, we conducted experiments comparing different model configurations: 1) TA-RIR(R): Eliminated the position topology encoding module from TA-RIR. Instead, we concatenated the room vertex, sound source, and receiver coordinates with the reverberated speech for embedding. 2) TA-RIR(G): Removed the reverberated speech encoder. By evaluating the performance differences among these configurations, the complete model, and the baseline FiNS [12], we determined the impact of the position topology encoding module.

The scarcity of real-world room impulse response datasets, primarily recorded in one or a few rooms, limits the number of available room information and RIRs, which is insufficient for comprehensive model training [19, 20, 21]. To address this limitation, we utilized the Geometric-Wave Acoustic (GWA) [22] framework to generate 20,000 synthetic RIRs for training purposes. GWA is a framework that simulates RIRs by integrating the geometric acoustic simulator Gsound [23] with a wave

Table 1: *Test performance on objective metrics across different models.*

Dataset	**Model**	**Loss** $\downarrow$	RT_{60}			CTE		
			$\lvert Bias\rvert \downarrow$	$\lvert MSE(s)\rvert \downarrow$	$\rho \uparrow$	$\lvert Bias\rvert \downarrow$	$\lvert MSE(s)\rvert \downarrow$	$\rho \uparrow$
GWA	TA-RIR(Ours)	1.237	0.033	0.0096	0.958	0.329	5.144	0.953
	TA-RIR(R)	1.361	0.053	0.0157	0.937	0.483	7.001	0.933
	TA-RIR(G)	2.330	0.268	0.174	0.146	1.287	58.88	0.146
	FiNS	1.543	0.055	0.0200	0.922	1.434	16.689	0.860

acoustic simulator based on the finite difference time domain (FDTD) method [24]. This integration enables the generation of high-quality synthetic RIRs that accurately represent various acoustic environments. We convolved these generated RIRs with clean speech samples from the VCTK [25] dataset to produce 200,000 reverberant speech samples. Both the RIRs and clean speech samples were processed at a sampling rate of 48 kHz. Each RIR was adjusted to a length of 48,000 samples (equivalent to 1 second), while each clean speech sample was cropped to 131,072 samples (approximately 2.73 seconds). For all experiments, the model was trained using the AdamW optimizer with an initial learning rate of 3.5×10^{-5}. The batch size was set to 64, and the learning rate was decayed by a factor of 0.8 every 80 epochs.

3.2. Embedding Analysis

We use UMAP [26] to perform a two-dimensional visualization of the encoder-generated fusion embedding vectors, with the colors representing the ground truth's RT_{60}value and CTE values. As shown in the figure 3, the projection of the embeddings in the two-dimensional space reveals a discernible distribution pattern. Samples show a relatively concentrated distribution within the feature space, with distinct separations observed between areas corresponding to different value ranges. This distribution structure indicates that the depth encoder successfully captured essential features related to the target index during training, validating the formation of implicit semantic structures throughout the characterization learning process.

3.3. Evaluation Results

To assess the quality of room impulse response estimates produced by each model, we employed several evaluation metrics: multi-resolution Short-Time Fourier Transform (STFT) error, RT_{60}, and CTE. RT_{60} represents the duration required for the sound pressure level to decay by 60 decibels. CTE quantifies the proportion of total sound energy received within the first 50 milliseconds relative to the energy received during the subsequent period throughout the entire sound reception process. For both RT_{60} and CTE, this study evaluated bias and mean squared error (MSE). The detailed results are presented in Table 1, which clearly demonstrate that our model outperforms the baseline in all evaluated metrics. Furthermore, the effectiveness of the proposed position topology encoding module and its integration with reverb speech encoding embedding are evident, as they contribute significantly to the improved performance observed.

4. Conclusion

This paper presents a method for utilizing position topology information to estimate time-domain impulse responses (RIRs) for reverberant signals. The approach employs graph neural networks to encode the positional relationships between the sound source, receiver, and room, and integrates these with embeddings extracted from reverberant speech to enhance RIR estimation performance. Experimental results demonstrate that the proposed positional encoding module enables the model to extract compact acoustic embeddings that effectively capture the underlying properties and spatial information of the RIR. Compared to the baseline, the model significantly improves the accuracy of RIR estimation. Future research directions include exploring data augmentation techniques to increase the robustness of the model, integrating physical models to better simulate acoustic environments, and investigating other strategies to further improve generalization across different room configurations and reverberation conditions.

5. Acknowledgement

This work was supported by the Program of National Development and Reform Commission under Grant No. 2404-340161-04-04-420064 and the National Natural Science Foundation of China under Grant No. 62171427.

6. References

[1] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2021.

[2] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[3] B. S. Liang, A. S. Liang, I. Roman, T. Weiss, B. Duinkharjav, J. P. Bello, and Q. Sun, "Reconstructing room scales with a single sound for augmented reality displays," *Journal of Information Display*, vol. 24, no. 1, pp. 1–12, 2023.

[4] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio engineering society*, vol. 50, no. 4, pp. 249–262, 2002.

[5] R. Falcon Perez, "Machine-learning-based estimation of room acoustic parameters," 2018.

[6] S. Sakamoto, A. Ushiyama, and H. Nagatomo, "Numerical analysis of sound propagation in rooms using the finite difference time domain method," *The Journal of the Acoustical Society of America*, vol. 120, no. 5_Supplement, pp. 3008–3008, 2006.

[7] A. Krokstad, S. Strom, and S. Sørsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, vol. 8, no. 1, pp. 118–125, 1968. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022460X68901983

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

[9] 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, Apr. 2022.

[10] A. Ratnarajah, Z. Tang, and D. Manocha, "TS-RIR: Translated Synthetic Room Impulse Responses for Speech Augmentation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Cartagena, Colombia: IEEE, Dec. 2021, pp. 259–266.

[11] I. Martin, F. Pastor, F. Fuentes-Hurtado, J. A. Belloch, L. Azpicueta-Ruiz, V. Naranjo, and G. Piñero, "Predicting room impulse responses through encoder-decoder convolutional neural networks," in *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2023, pp. 1–6.

[12] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 221–225.

[13] Z. Liao, F. Xiong, J. Luo, M. Cai, E. S. Chng, J. Feng, and X. Zhong, "Blind estimation of room impulse response from monaural reverberant speech with segmental generative neural network." INTERSPEECH, 2023.

[14] A. Ratnarajah, S. Ghosh, S. Kumar, P. Chiniya, and D. Manocha, "Av-rir: Audio-visual room impulse response estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 164–27 175.

[15] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "Film: visual reasoning with a general conditioning layer," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:13740328

[18] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204915831

[19] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černockỳ, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[20] A. Kujawski, A. J. Pelling, and E. Sarradj, "Miracle—a microphone array impulse response dataset for acoustic learning," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 32, 2024.

[21] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *2009 16th International Conference on Digital Signal Processing*. IEEE, 2009, pp. 1–5.

[22] Z. Tang, R. Aralikatti, A. J. Ratnarajah, and D. Manocha, "Gwa: A large high-quality acoustic dataset for audio processing," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–9.

[23] C. Schissler and D. Manocha, "Gsound: Interactive sound propagation for games," in *Audio Engineering Society Conference: 41st International Conference: Audio for Games*. Audio Engineering Society, 2011.

[24] B. Hamilton, "Pffdtd software," 2021, https://github.com/bsxfun/pffdtd.

[25] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:213060286

[26] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.