Interspeech 2024
1-5 September 2024, Kos, Greece

RIR-in-a-Box: Estimating Room Acoustics from 3D Mesh Data through Shoebox Approximation

Liam Kelley[1], *Diego Di Carlo*[2], *Aditya Arie Nugraha*[2],
Mathieu Fontaine[3,2], *Yoshiaki Bando*[2], *Kazuyoshi Yoshii*[4,2]

[1] ENSTA Paris, IP Paris, France, [2] Center for Advanced Intelligence Project, RIKEN, Japan,
[3] LTCI, Télécom Paris, IP Paris, France, [4] Graduate School of Engineering, Kyoto University, Japan,
liam.kelley@ensta-paris.fr, yoshii.kazuyoshi.3r@kyoto-u.ac.jp

Abstract

This paper describes a method for estimating the room impulse response (RIR) for a microphone and a sound source located at arbitrary positions from the 3D mesh data of the room. Simulating realistic RIRs with pure physics-driven methods often fails the balance between physical consistency and computational efficiency, hindering application to real-time speech processing. Alternatively, one can use MESH2IR, a fast black-box estimator that consists of an encoder extracting latent code from mesh data with a graph convolutional network (GCN) and a decoder generating the RIR from the latent code. Combining these two approaches, we propose a fast yet physically coherent estimator with interpretable latent code based on differentiable digital signal processing (DDSP). Specifically, the encoder estimates a virtual shoebox room scene that acoustically approximates the real scene, accelerating physical simulation with the differentiable image-source model in the decoder. Our experiments showed that our method outperformed MESH2IR for real mesh data obtained with the depth scanner of Microsoft HoloLens 2, and can provide correct spatial consistency for binaural RIRs.
Index Terms: Spatial audio, room acoustics, 3D mesh data, physical models, DDSP

1. Introduction

The room impulse response (RIR) serves as a cornerstone in a wide variety of audio signal processing systems, from speech enhancement for automatic speech recognition [1] to auralization of dry signals for spatial synthesis [2]. It encodes the propagation of sound waves in an indoor environment and is affected by the geometry of the room, the materials lining its surfaces, and the positions of sources and microphones within the space [3]. Efficient RIR simulation plays a vital role in virtual/mixed reality systems for enhancing the immersive experience [4, 5]. It must be able to quickly adapt to highly dynamic environments at real time, aligning with the visual information of both virtual and real spaces. RIR simulation has also been used for making realistic speech datasets expected to improve the robustness of speech analysis models in various environments [6, 7].

In addition to datasets of RIRs measured in real rooms, one may artificially generate realistic RIRs using a physics-based acoustic simulator, such as the image-source method (ISM) [8], which aims to find the purely specular reflection paths between a source and a microphone, and its accelerated variants [9–11]. Unfortunately, some acceleration techniques, even ones with physically motivated assumptions, may violate certain physical properties that affect the representation of early reflections crucial for maintaining the spatial cues [7, 12].

To bypass computationally intensive physics-based simulation, RIR estimators based on deep learning have recently been developed [13–15]. These methods tend to work fast at run-time and are capable of fusing multimodal clues related to the RIR. In addition to audio signals recorded by microphones, visual information such as raw images [16–19] have been effectively used for RIR estimation. In particular, we focus on MESH2IR [15] that estimates the RIR from the 3D mesh data of the room because it is useful for audio rendering in mixed reality applications using LiDAR scanners (e.g., smart glasses and MR headsets). It is based on the encoder-decoder architecture; the encoder embeds the 3D mesh data along with source and microphone positions to the latent code with a graph convolutional network (GCN) and the decoder generates the RIR from the latent code. However, the physical consistency of the estimated RIR is not guaranteed due to the data-driven nature of the decoder.

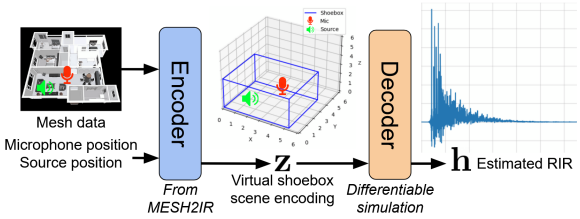

Figure 1: *The overview of RIR-in-a-Box, a novel physics-aware mesh-to-RIR translator trained such that the RIR simulated in the virtual shoebox scene (explainable latent code) is made close to the RIR measured in the real complicated scene.*

To leverage the complementary advantages of the physics-based simulator and the data-driven estimator, we propose "RIR-in-a-Box", an ISM-injected version of MESH2IR based on differentiable digital signal processing (DDSP) [20]. As shown in Fig. 1, the encoder transforms the 3D mesh data into an *explainable* latent vector representing a virtual shoebox room scene that acoustically approximates the target room scene. This enables the decoder to efficiently simulate the RIR based on the ISM for the estimated shoebox room thanks to its simple shape, even if the ISM-based simulation for the original mesh data is computationally intensive. Given the ground-truth RIR, the whole network can be trained by backpropagating the loss to the encoder through the decoder, the *differentiable* ISM.

The main contribution of this study is to develop a DDSP-based method that can reliably simulate the RIR by approximating the target scene as a shoebox scene. Another contribution to the field is to provide a real mesh-audio dataset obtained with the microphones and LiDAR scanner of Microsoft HoloLens 2 and a Pytorch implementation of the differentiable ISM.

2. Related work

This section reviews related work in terms of individual techniques used in the proposed method.

10.21437/Interspeech.2024-2053

Table 1: *RIR-in-a-Box vs baselines.*

Model	Interpret.	Physical	Material-blind	Fast
ISM (in a Shoebox)	✓	✓	×	✓
FASTRIR [21]	✓	×	×	✓
GWA [22]	✓	✓	×	×
MESH2IR [15]	×	×	✓	✓
RIRBox (proposed)	✓	✓	✓	✓

2.1. Acoustic matching

Acoustic matching techniques aim to reflect the acoustics of a target environment in audio recordings. This field includes estimating high-level acoustic parameters such as the direct-to-reverberant ratio (DRR) and reverberation time (RT60) to the actual sound field through RIRs (see [17, Sec. 2] for a review). Recent advancements have expanded the scope of acoustic matching, incorporating audio embeddings and even visual elements to define target acoustic spaces [15, 17]. Notably, the integration of visual information within deep learning frameworks has opened new avenues for acoustic matching, enabling the estimation of RIRs through images [19] and leveraging machine learning models for enhanced spatial audio realism.

2.2. Acoustic simulation

Acoustic simulators can be roughly categorized into two groups: physics-based and deep generative models (Table 1). Physics-based simulators (e.g., [22]) use computationally intensive software to model sound propagation. Simplified room acoustic models (e.g. the ISM) reduce the computation load by retaining certain physical and perceptual properties. Thanks to this balance, they are the go-to methods in most applications. Nevertheless, such simulators typically require prior information (e.g., the RT60, the room shape, and wall absorption coefficients), which may be challenging to obtain in practice.

Thanks to their fast inference and ability to extract knowledge from data, deep generative models (e.g., FASTRIR [21] and MESH2IR [15]) have been proposed for RIR generation. FASTRIR is designed for shoebox configurations with known RT60, room dimensions, and microphone and source positions. MESH2IR extends it to complex room layouts by relying on a black-box mesh deep encoder. Interestingly, this approach demonstrates that acoustic properties (e.g., wall absorption or RT60) could be inferred blindly from the room mesh. Both algorithms are fast during inference compared to the physics-based acoustic simulators, but function as opaque, data-driven systems.

To address this issue, we introduce `RIRBox`, a tool capable of inferring the acoustic properties of a room from its detailed mesh, which can nowadays be easily captured by head-mounted displays (e.g., Hololens2). `RIRBox` generates RIRs based on the ISM, aiming to be faster than advanced physics-based simulators while providing a balance between interpretability and computational efficiency.

2.3. Differentiable digital signal processing

DDSP enables us to jointly optimize classical acoustic models in tandem with neural networks. This approach has been recently extended to model reverberation [13, 23, 24]. Besides performance improvements with respect to pure data-driven methods, these models also gain interpretability. Interestingly, the work of [24] uses a differentiable implementation of the ISM to solve acoustic inverse problems. Our work builds upon this innovative intersection, utilizing the ISM as a physics-constrained decoder within the DDSP framework.

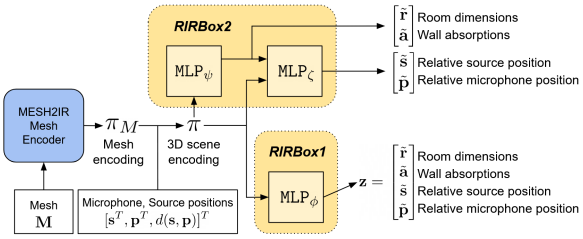

Figure 2: *Encoder architecture.*

3. Proposed method

This section explains the proposed fast yet physically consistent RIR estimator, `RIRBox`.

3.1. Problem Specification

Given the 3D triangular mesh data of a target room $\mathbf{M}$, we aim to estimate the time-domain RIR [15] $\mathbf{h} \in \mathbb{R}^L$ with respect to a source position $\mathbf{s} \in \mathbb{R}^3$ and a microphone position $\mathbf{p} \in \mathbb{R}^3$ represented in the Cartesian coordinate system as follows:

$$\mathbf{h} = \texttt{RIRBox}(\mathbf{M}, \mathbf{s}, \mathbf{p}), \tag{1}$$

where L represents the length of the RIR to be considered. The mesh data $\mathbf{M}$ is given as a graph represented as an adjacency matrix over nodes.

3.2. Model design

`RIRBox` is a DDSP-based estimator with the encoder-decoder architecture consisting of the encoder of MESH2IR [15] and a decoder based on the differentiable ISM [24].

$$\mathbf{z} = \texttt{Encoder}(\mathbf{M}, \mathbf{s}, \mathbf{p}), \tag{2}$$

$$\mathbf{h} = \texttt{Decoder}(\mathbf{z}), \tag{3}$$

where $\mathbf{z}$ is an explainable latent vector representing a scene of a shoebox room as follows:

$$\mathbf{z} \triangleq \left[\tilde{\mathbf{r}}^\mathsf{T}, \tilde{\mathbf{a}}^\mathsf{T}, \tilde{\mathbf{s}}^\mathsf{T}, \tilde{\mathbf{p}}^\mathsf{T}\right]^\mathsf{T} \in \mathbb{R}^{12}, \tag{4}$$

where $\tilde{\mathbf{r}} \in \mathbb{R}^3_{\geq 1}$ represents the dimensions of the virtual shoebox room, $\tilde{\mathbf{a}} \in [0.01, 0.85]^3$ represents the absorption coefficients of the walls, the floor, and the ceiling, and $\tilde{\mathbf{s}}, \tilde{\mathbf{p}} \in [0, 1]^3$ represent the positions of the source and microphone normalized with respect to the room dimensions. Note that $\mathbf{s}$ and $\mathbf{p}$ are the positions in the original 3D scene, while $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{p}}$ are those in the equivalent virtual shoebox scene.

The key feature of `RIRBox` is that it estimates the virtual shoebox scene $\mathbf{z}$ in which the RIR $\mathbf{h}$ can be simulated efficiently in a physically faithful manner, instead of directly estimating $\mathbf{h}$ with a blackbox network as in MESH2IR [15]. This calls for the physics-based decoder implemented as the differentiable ISM [15] and enables the supervised training of the whole network with backpropagation. The ISM-based simulation for shoebox rooms is much faster than that for real rooms represented by complicated mesh data.

3.2.1. Encoder

As shown in Fig. 2, the encoder internally estimates an intermediate representation $\boldsymbol{\pi}$ required for constructing the virtual shoebox scene, which is given by

$$\boldsymbol{\pi} \triangleq \left[\boldsymbol{\pi}_\mathbf{M}^\mathsf{T}, \mathbf{s}^\mathsf{T}, \mathbf{p}^\mathsf{T}, d(\mathbf{s},\mathbf{p})\right]^\mathsf{T} \in [0, 1]^{15}, \tag{5}$$

where $d(\mathbf{s}, \mathbf{p})$ is the distance between the source and microphone and $\boldsymbol{\pi}_\mathbf{M} \in [0, 1]^8$ is a latent vector extracted from the mesh data $\mathbf{M}$ with the encoder of MESH2IR based on a graph convolutional

network (GCN) [15] as follows:

$$\boldsymbol{\pi}_{\mathbf{M}} = \texttt{GCN}(\mathbf{M}) \in [0,1]^8. \tag{6}$$

We propose two variants of the proposed method. `RIRBox1` computes the latent vector $\mathbf{z}$ directly from the intermediate representation $\boldsymbol{\pi}$ as follows:

$$\mathbf{z} = \texttt{MLP}_\phi(\boldsymbol{\pi}), \tag{7}$$

where $\texttt{MLP}_\phi$ is a multilayer perceptron network with parameters ϕ. In contrast, `RIRBox2` takes two steps; It first estimates the dimensions of a shoebox room $\tilde{\mathbf{r}}$ and its absorption coefficients $\tilde{\mathbf{a}}$, which are then used to condition the estimation of the source position $\tilde{\mathbf{s}}$ and microphone position $\tilde{\mathbf{p}}$ as follows:

$$\left[\tilde{\mathbf{r}}^\mathsf{T},\, \tilde{\mathbf{a}}^\mathsf{T}\right]^\mathsf{T} = \texttt{MLP}_\psi(\boldsymbol{\pi}), \tag{8}$$

$$\left[\tilde{\mathbf{s}}^\mathsf{T},\, \tilde{\mathbf{p}}^\mathsf{T}\right]^\mathsf{T} = \texttt{MLP}_\zeta\left(\boldsymbol{\pi},\, \tilde{\mathbf{r}},\, \tilde{\mathbf{a}}\right), \tag{9}$$

where $\texttt{MLP}_\psi$ and $\texttt{MLP}_\zeta$ are separate sub-networks parameterized by ψ and ζ, respectively.

3.2.2. Decoder

The decoder generates the RIR corresponding to the virtual shoebox scene with the classical yet physics-faithful ISM-based simulation. Using the idea of DDSP [20], the ISM is implemented in a differentiable manner, making it possible to train the encoder in a supervised manner by backpropagating the estimation error through the decoder. Since the ISM suffers from artifacts under 80 Hz for shoebox rooms [6], we finally apply a high-pass filter with a cutoff frequency of 80 Hz for the RIR obtained with the ISM as follows:

$$\mathbf{h} \leftarrow \texttt{HP}_{\texttt{80Hz}}(\texttt{ISM}(\mathbf{z})). \tag{10}$$

3.3. Parameter optimization

We optimize the network parameters, ϕ of `RIRBox1` or ψ and ζ of `RIRBox2`, such that the estimated RIR $\mathbf{h}$ well approximates the ground-truth RIR $\hat{\mathbf{h}}$ for the given scene in the real room. In this paper, we propose three loss functions that evaluate the RIR estimation errors.

Multi-resolution short-time Fourier transform (STFT) loss represents the distance between the magnitude spectrograms of the estimated and ground-truth RIRs. It is obtained by averaging the sums of the spectral convergence losses and the log STFT magnitude losses over multiple time-frequency configurations of STFT [25].

Energy decay relief (EDR) loss $\mathcal{L}_{\text{EDR}}$ represents the distance between the EDR of the estimated and ground-truth RIRs. The energy decay curve (EDC) represents how the total energy left in a sound signal decreases smoothly over time, which is crucial for understanding the reverberation time of a room. The EDR is the generalized version of the EDC for multiple frequency bands [15].

Distance loss $\mathcal{L}_{\mathbf{d}}$ represents the distance between the estimated microphone-source distance in the virtual shoebox room and the ground-truth distance in the real room:

$$\mathcal{L}_{\text{d}}(\mathbf{s}, \mathbf{p}, \tilde{\mathbf{r}}, \tilde{\mathbf{s}}, \tilde{\mathbf{p}}) \triangleq \|d(\mathbf{s}, \mathbf{p}) - d(\tilde{\mathbf{r}} \odot \tilde{\mathbf{s}}, \tilde{\mathbf{r}} \odot \tilde{\mathbf{p}})\|_2^2, \tag{11}$$

where $\odot$ denotes the element-wise multiplication.

The total loss $\mathcal{L}$ is given by

$$\mathcal{L} \triangleq \mathcal{L}_{\text{MRSTFT}} + \lambda_1 \mathcal{L}_{\text{EDR}} + \lambda_2 \mathcal{L}_{\text{d}}, \tag{12}$$

where $\lambda_1, \lambda_2 \in \mathbb{R}_+$ are the weighting coefficients of $\mathcal{L}_{\text{EDR}}$ and $\mathcal{L}_{\text{d}}$. The whole network is trained in a supervised manner such that the total loss function $\mathcal{L}$ is minimized.

4. Evaluation

This section reports experiments conducted to evaluate the RIRs generated by the proposed method in terms of accuracy and spatial consistency.

4.1. Experimental conditions

The baseline models, including MESH2IR [15], were trained on *synthetic* RIRs from the GWA dataset [22] and tested on separate synthetic RIRs and *real* RIRs recorded using a Microsoft HoloLens 2 (HL2) in real environments[1].

4.1.1. Synthetic dataset

The proposed model was trained on a subset of the GWA dataset [22], which originally includes 2 million synthetic RIRs computed with a high-quality hybrid wave-geometric simulation in 6.8k hand-crafted 3D rooms of the 3D-FRONT dataset [26]. This training subset contained 3D scenes where a direct path is likely to be present between the microphone and the source since that situation is the one our shoebox model will best represent. In those scenes, the first peak of the computed IR was within 20 samples of the onset of an unblocked direct path and the first peak amplitude was within 30% of the amplitude of an unblocked direct path. This training subset contained 9005 RIRs (scenes) in 4297 rooms. Using the same procedure, we also created a test subset that contains 520 RIRs (scenes) in 236 rooms.

Following [22], we downsampled the RIRs to 16 kHz and used only the first 3968 samples. Although, since our physics-based model innately models the energy, we did not perform the standard deviation normalization and, in contrast, tried to recover the correct RIR scale from the downscaled audio files in the GWA dataset. We also performed mesh simplification preprocessing, as in [22], to have a consistent number of mesh faces (i.e., 2000) from any input with a varying number of faces.

4.1.2. Real dataset

In addition to the synthetic test dataset above, we also constructed a real test dataset consisting of 481 scenes in total. It includes 30 room meshes of two connecting rooms: a meeting room with a size of 6.3×3.5×3.1 m and a small office room with a size of 2.5×3.5×3.1 m. Both rooms have walls that are relatively reflective due to whiteboards, windows, or flat concrete walls. The floors are covered with carpet, while the ceilings have lighting and air conditioning installations. The dataset also includes 27 RIRs from different combinations of microphone and source positions in those rooms.

The RIRs were estimated from real audio signals of exponential sine sweeps [27] emitted by a high-fidelity directional loudspeaker (Genelec 8030C) and recorded by HL2 worn on a human head. The source and microphone positions were manually measured, so they are approximate. The room meshes were also captured using HL2 worn on a human head by utilizing the functionalities of the Windows Device Portal for HL2. In practice, the quality of the mesh increases in terms of accuracy and coverage as the user explores the room. We captured this effect by iteratively acquiring the room mesh following three different protocols: a progressive scan while sitting down in the meeting room and looking around, a progressive scan walking throughout the meeting room first and then into the small office, and a progressive scan walking throughout both rooms at once. We applied the same mesh simplification preprocessing as for the training dataset.

[1] Code and data: https://github.com/liam-kelley/RIR-in-a-Box

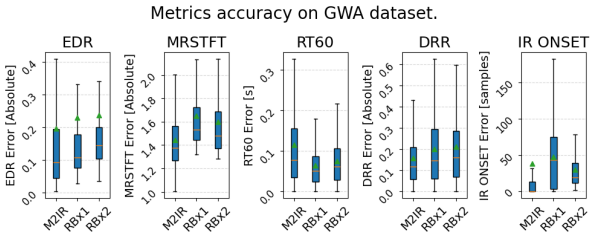

Figure 3: *Metric performance on GWA held-out validation set.*

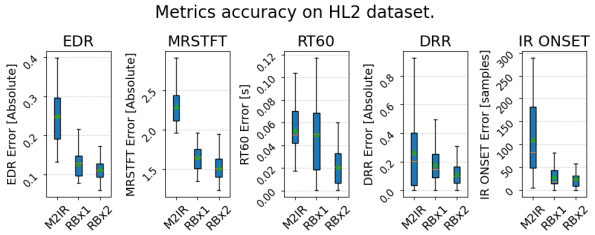

Figure 4: *Metric performance on HL2 validation set.*

Table 2: *Average AOA errors with the standard deviations.*

MESH2IR	RIRBox1	RIRBox2
$45.1° \pm 26.1°$	$36.0° \pm 22.6°$	$\mathbf{24.8°} \pm \mathbf{19.3°}$

4.2. Implementation details

`RIRBox1` and `RIRBox2` used the pre-trained mesh encoder of MESH2IR [15], and $\texttt{MLP}_\phi$, $\texttt{MLP}_\psi$, and $\texttt{MLP}_\zeta$ had three hidden layers with rectified linear units (ReLU) and an output layer that applies the softplus activation function for $\tilde{\mathbf{r}}$, and the sigmoid activation function for $\tilde{\mathbf{p}}$, $\tilde{\mathbf{s}}$, and $\tilde{\mathbf{a}}$. The decoder used the differentiable ISM [24] with a lower order of reflections at training time (10) than at test time (18). In preliminary experiments, we observed that during training the benefits of a larger batch size outweighed the inaccuracies of having fewer reflections. Our models were trained on the synthetic dataset (Section 4.1.1) with a batch size of 28 for 12 epochs using the Adam optimizer with an initial learning rate of 10^{-3} scheduled to decrease 70% every 4 epochs. The weighting coefficients were set to $\lambda_1 = 0.5$ and $\lambda_2 = 2.0$.

4.3. Experimental results

We discuss the performance of the proposed method in the RIR estimation and the downstream task.

4.3.1. RIR estimation

We compared `RIRBox` with MESH2IR [15] on the GWA held-out-validation set and the real-life HoloLens2 validation dataset in terms of EDR, MRSTFT, Reverberation Time (RT60), Direct to Reverberant Ratio (DRR), and difference in RIR onset time (IR Onset). As shown in Figs. 3 and 4, `RIRBox2` was more robust in the HL2 validation set, and `RIRBox1` estimated RT60 best on the GWA held-out validation set. An example inference on the HL2 validation set is shown in Fig. 5. MESH2IR's RIR is very degraded on real data, while `RIRBox2` is more robust, even though they use the same latent mesh encoding π_M. The virtual room dimensions inferred by `RIRBox2` ($4.0\times2.5\times1.9$m) are off by about 35% from the meeting room dimensions ($6.3\times3.5\times3.1$m) although proportions are consistent, and the virtual mic-source distance 1.9m is off by 13.5% from the ground truth 2.2m.

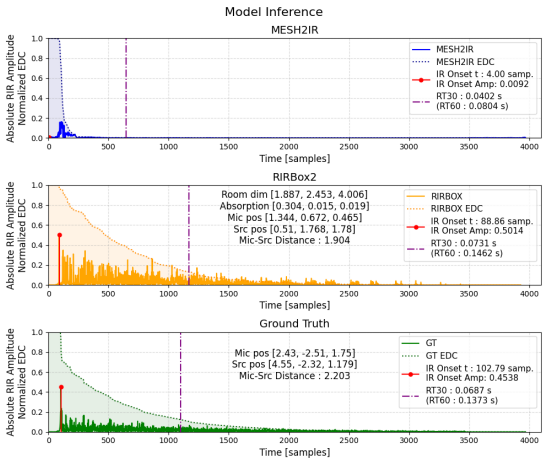

Figure 5: *Example of RIR generation for a HL2-sampled mesh.*

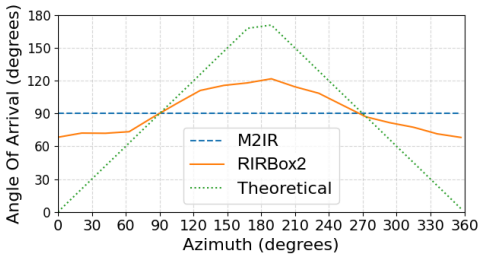

Figure 6: *Example AOA vs azimuth for a GWA validation mesh.*

4.3.2. RIR spatial consistency

By calculating the time difference of arrival of sound for 2 microphones imitating an HL2 headset (at a distance of 22.5 cm), we can estimate the angle of arrival (AOA) of a sound source. We compared the MESH2IR [15] and `RIRBox` methods' generated RIR spatial consistency through their AOA estimation performance. We used the GWA held-out-validation set meshes, put the microphone array at a random position within each mesh, and simulated sound originating from different azimuth angles. We computed the AOA for sound reaching the microphones using GCC-PHAT [28]. We compared the results with the theoretical AOA based on the geometry of the setup. The performance of each model against theoretical values is listed in Table 2. Fig. 6 showed that MESH2IR [15] outputs a constant AOA. Further inspection showed that MESH2IR's RIRs for the two microphones slightly differ in amplitudes but usually not in the direct path location, resulting in a constant AOA. Meanwhile, `RIRBox` struggles with extreme angles but accurately maintains spatial cues near the frontal direction.

5. Conclusion

We have presented a DDSP-based method to estimate room impulse responses (RIRs) from complex 3D scenes. Specifically, we extended the existing neural estimator MESH2IR to produce more physically consistent RIRs by incorporating a differentiable implementation of the image source model. Experimental results on real mesh data given by Microsoft HoloLens 2 showed that the proposed method outperformed MESH2IR in both RIR estimation and RIR spatial consistency. In future work, we aim to apply our method to other downstream tasks in mixed reality applications, e.g., speech enhancement and dereverberation. Including a statistical model for late reverberation, along with training the mesh encoder end-to-end will also be investigated.

6. Acknowledgment

This work was supported by ANR Project SAROUMANE (ANR-22-CE23-0011) and Hi! Paris Project MASTER-AI, JST PRESTO no. JPMJPR20CB, and JSPS KAKENHI nos. JP20H00602, JP21H03572, JP23K16912, JP23K16913.

7. References

[1] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[2] S. Liu and D. Manocha, *Sound Synthesis, Propagation, and Rendering*. Springer, 2022.

[3] H. Kuttruff, *Room Acoustics*, 6th ed. CRC Press, 2016.

[4] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher, "Virtual reality system with integrated sound field simulation and reproduction," *EURASIP J. Adv. Signal Process.*, vol. 2007, pp. 1–19, 2007.

[5] S. Serafin, M. Geronazzo, C. Erkut, N. C. Nilsson, and R. Nordahl, "Sonic interactions in virtual reality: State of the art, current challenges, and future directions," *IEEE Comput. Graph. Appl.*, vol. 38, no. 2, pp. 31–43, 2018.

[6] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," in *Proc. INTERSPEECH*, 2017, pp. 379–383.

[7] Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha, "Improving reverberant speech training using diffuse acoustic simulation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6969–6973.

[8] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Amer. Stat. Assoc.*, vol. 65, no. 4, pp. 943–950, 1979.

[9] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *J. Amer. Stat. Assoc.*, vol. 138, no. 2, pp. 708–730, 2015.

[10] Z.-H. Fu and J.-W. Li, "GPU-based image method for room impulse response calculation," *Multimed. Tools Appl.*, vol. 75, pp. 5205–5221, 2016.

[11] Y. Luo and J. Yu, "FRA-RIR: Fast random approximation of the image-source method," in *Proc. INTERSPEECH*, 2023, pp. 3884–3888.

[12] A. Aknin, T. Dupré, and R. Badeau, "Evaluation of a stochastic reverberation model based on the image source principle," in *Proc. Int. Conf. Digital Audio Effects*, 2020, pp. 31–37.

[13] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2021, pp. 221–225.

[14] S. Lee, H.-S. Choi, and K. Lee, "Yet another generative model for room impulse response estimation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023, pp. 1–5.

[15] A. Ratnarajah, Z. Tang, R. Aralikatti, and D. Manocha, "MESH2IR: Neural acoustic impulse response generator for complex 3D scenes," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 924–933.

[16] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori, "Image2Reverb: Cross-modal reverb impulse response synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2021, pp. 286–295.

[17] C. Chen, R. Gao, P. Calamia, and K. Grauman, "Visual acoustic matching," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognition*, 2022, pp. 18 858–18 868.

[18] B. Ahn, K. Yang, B. Hamilton, J. Sheaffer, A. Ranjan, M. Sarabia, O. Tuzel, and J.-H. R. Chang, "Novel-view acoustic synthesis from 3D reconstructed rooms," 2023, arXiv:2310.15130v1.

[19] A. Ratnarajah, S. Ghosh, S. Kumar, P. Chiniya, and D. Manocha, "AV-RIR: Audio-visual room impulse response estimation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognition*, 2024, pp. 27 164–27 175.

[20] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–12.

[21] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "FAST-RIR: Fast neural diffuse room impulse response generator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 571–575.

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

[23] S. Lee, H.-S. Choi, and K. Lee, "Differentiable artificial reverberation," *IEEE/ACM Trans. ASLP*, vol. 30, pp. 2541–2556, 2022.

[24] B. Zhi, A. Sharma, D. N. Zotkin, and R. Duraiswami, "A differentiable image source model for room acoustics optimization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023, pp. 1–5.

[25] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6199–6203.

[26] H. Fu, B. Cai, L. Gao, L.-X. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao *et al.*, "3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2021, pp. 10 933–10 942.

[27] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Proc. Audio Eng. Soc. Conv.*, 2007, pp. 1–21.

[28] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. ASSP*, vol. 24, no. 4, pp. 320–327, 1976.