MESH2IR: Neural Acoustic Impulse Response Generator for Complex 3D Scenes

Anton Ratnarajah, Zhenyu Tang, Rohith Aralikatti, Dinesh Manocha

University of Maryland, College Park, MD, USA

Paper ID : mmfp2106

ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

Motivation

- Rapid development in interactive applications (e.g., games and virtual environments) demands realistic sound effects in a complex indoor environment with multiple sources.
- No current simulation and learning methods can compute real-time IRs for unseen complex dynamic scenes.

Main Contributions

- We propose a novel learning-based IR generator (MESH2IR) to generate realistic IRs for furnished 3D scenes with arbitrary topologies in real-time.
- We present an efficient approach to preprocess the IR training dataset and show that training MESH2IR on preprocessed dataset gives a significant improvement in the accuracy of IR generations.
- Our MESH2IR can generate IRs 200 times faster than a geometric acoustic simulator on a single CPU.
- We also show that far-field speech augmented using the IRs generated from MESH2IR significantly improves the performance in speech processing applications.

Architecture

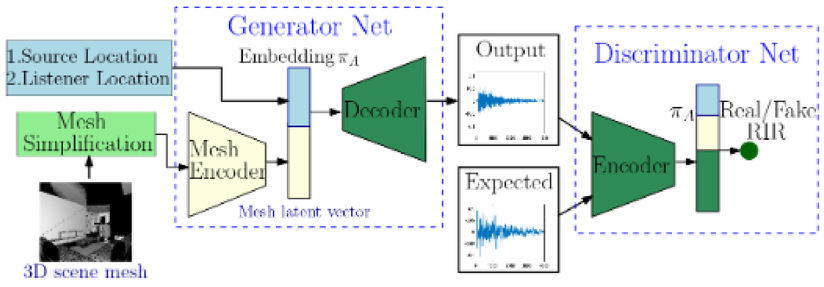
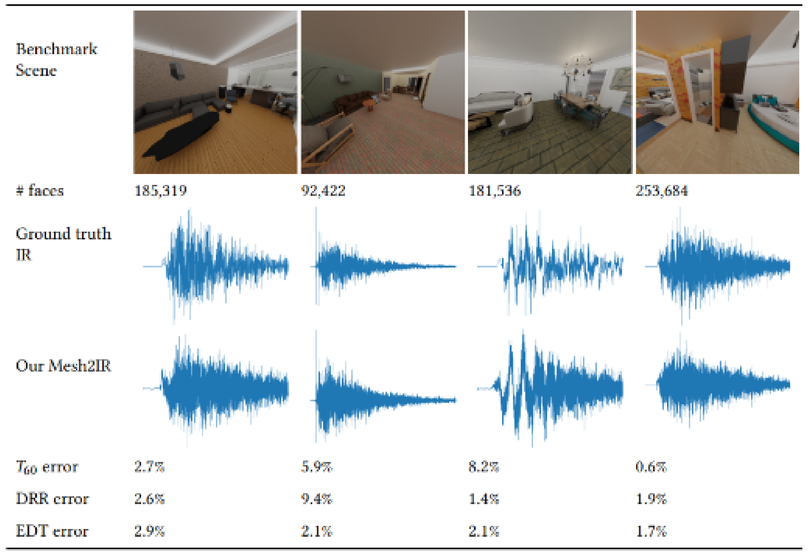


Figure 1: The architecture of our MESH2IR. Our mesh encoder network encodes an indoor 3D scene mesh to the latent space. The mesh latent vector and source and listener positions are combined to produce a scene vector embedding (π_A). The generator network generates IR corresponding to the input scene vector embedding. For the given scene vector embedding, the discriminator network discriminates between the generated IR and the ground truth IR during training.

Acoustic Evaluation



Ablation Experiments

- To evaluate the importance and the efficient way of adding Energy Decay Relief (EDR) to the cost function, we train and compare 2 different variations of our MESH2IR network.
  - **Variation 1 (MESH2IR-NO-EDR):** MESH2IR without the EDR loss.
  - **Variation 2 (MESH2IR-D-EDR):** We train the Discriminator network to discriminate EDR of the generated IRs and the ground truth IRs.
- EDR is the total amount of energy remaining in the IR at time t_n seconds in a frequency band centered at f_k Hz.

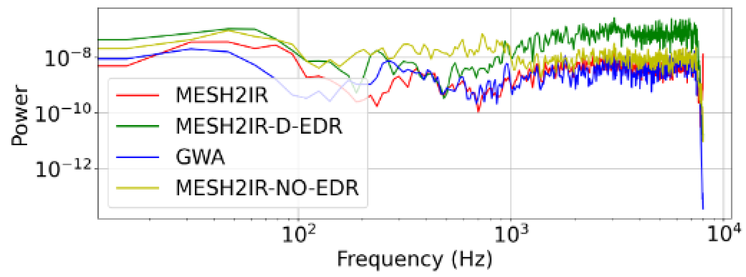


Figure 2: The power spectrum of IRs. We can see that the power spectrum of MESH2IR is closest to the power spectrum of GWA.

Applications

- For a fair comparison between different methods, we generate 11K IRs from 600 scene meshes not used during training our MESH2IR using GWA, GA, MESH2IR-D-EDR and MESH2IR.
- We generate reverberant speech using the 11K IRs, and train **speech dereverberation** and s**peech separation methods**.

Table 1: Speech dereverberation results are obtained when training data is generated by different synthetic IR generation methods. Testing is done on reverberant data synthesized from IRs present in three different datasets containing recorded IRs collected in a variety of environments. Higher SRMR is better.

Training Dataset	SRMR		
	MIT	BUTReverb	RWCP Aachen
Reverb	7.35	3.14	5.16
GA [2]	6.39	3.74	4.83
GWA [1]	**7.67**	**4.6**	**6.14**
MESH2IR-D-EDR	6.18	3.29	4.32
MESH2IR (ours)	**7.82**	**4.27**	**5.88**

Table 2: {Speech separation results in the presence of reverberation are shown below. We report the improvement in the Scale-Invariant Signal Distortion Ratio (SI-SDRi) over the reverberant mixture. Higher Si-SDRi is better. We report performance on reverberant mixtures generated in four different room configurations present in the VOiCES dataset.

Training Dataset	SI-SDRi			
	Room 1	Room 2	Room 3	Room 4
GA [2]	2.26	2.22	1.33	2.35
GWA [1]	**4.75**	**4.75**	**2.41**	**4.91**
MESH2IR-D-EDR	4.68	4.35	1.87	4.72
MESH2IR (ours)	**4.91**	**4.89**	**2.54**	**5.13**

Runtime

Table 3: The runtime of a geometric acoustic simulator (GA), FAST-RIR, and our MESH2IR. MESH2IR is an extension of FAST-RIR to generate IRs for complex indoor scenes. We can see that the runtime of MESH2IR is higher than FAST-RIR because we use a graph-based network to process the mesh. MESH2IR still outperforms GA on a single CPU.

IR Generator	Hardware	Avg time	Scene Type
GA	CPU	30.05s	Simple
MESH2IR	**CPU**	**0.13s**	**Complex**
MESH2IR(Batch Size 1)	GPU	$1.32\text{x}10^{-2}$s	Complex
FAST-RIR(Batch Size 128)	GPU	$5.9\text{x}10^{-5}$s	Simple
MESH2IR(Batch Size 128)	**GPU**	$\mathbf{2.6\text{x}10^{-3}}$**s**	**Complex**
MESH2IR[Mesh Encoder]	**GPU**	$\mathbf{2.57\text{x}10^{-3}}$**s**	**Complex**
MESH2IR[IR Generator]	**GPU**	$\mathbf{7.4\text{x}10^{-5}}$**s**	**Complex**

Conclusion and Future Work

- We show that IRs predicted by our MESH2IR in unseen indoor 3D scenes are highly similar to the ground truth IRs generated from the GWA dataset, which is used to train our MESH2IR.
- The main limitation of our work is that we cannot control the characteristics of the scene materials such as the amount of sound absorption and scattering. Efficiently inputting the characteristics of scene material to our MESH2IR may improve the accuracy of IR generation.