FAST-RIR : FAST NEURAL DIFFUSE ROOM IMPULSE RESPONSE GENERATOR

Paper ID: 1736

Anton Ratnarajah[1], Shi-Xiong Zhang[2], Meng Yu[2], Zhenyu Tang[1], Dinesh Manocha[1], Dong Yu[2]

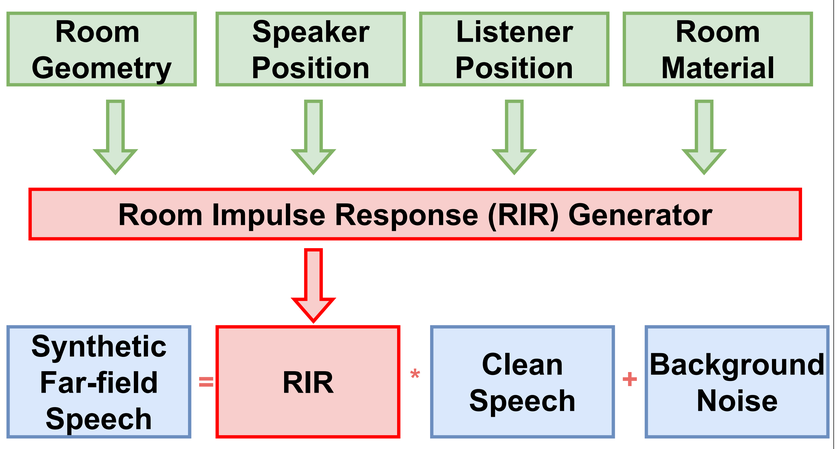
[1] University of Maryland, College Park, MD, USA

[2] Tencent AI Lab, Bellevue, WA, USA

Motivation

- With advancements in deep neural-network-based far-field speech processing the demand for **on-the-fly simulation** of far-field speech training datasets with **hundreds of thousands of room configurations similar to the testing environment** is increasing.
- We need a fast room impulse response (RIR) generator that can generate thousands of RIRs per second to simulate a large-scale far-field speech training dataset.

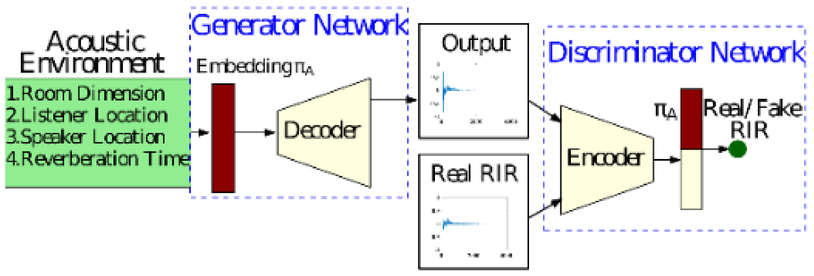
Introduction



Main Contributions

- We propose a neural-network-based fast diffuse room impulse response generator (FAST-RIR).
- Our FAST-RIR takes a constant amount of time to generate an RIR for any given acoustic environment, and yields accurate reverberation time.
- We trained our FAST-RIR to generate both specular and diffuse reflections of a given acoustic environment.
- Our FAST-RIR can generate RIRs 400 times faster than a state-of-the-art diffuse acoustic simulator (DAS) and performs similar to DAS in far-field automatic speech recognition (ASR) experiments.

Architecture



Experiments and Results

- We randomly select 30,000 different acoustic environments within the range of the training dataset and generate RIRs corresponding to the selected acoustic environments using the following RIR generators to evaluate the performance of FAST-RIR.

1. Image method
2. gpuRIR
3. Diffuse Acoustic Simulator (DAS)
4. **FAST-RIR**

Table 1. The runtime for generating 30,000 RIRs using image method, gpuRIR, DAS, and our FAST-RIR. Our FAST-RIR significantly outperforms all other methods in runtime.

RIR Generator	Hardware	Total Time	Avg Time
DAS [7]	CPU	$9.01\text{x}10^{5}$s	30.05s
Image Method [5]	CPU	$4.49\text{x}10^{3}$s	0.15s
FAST-RIR(Batch Size 1)	**CPU**	$\mathbf{2.15\text{x}10^{3}}$**s**	**0.07s**
gpuRIR [13]	GPU	16.63s	$5.5\text{x}10^{-4}$s
FAST-RIR(Batch Size 1)	GPU	34.12s	$1.1\text{x}10^{-3}$s
FAST-RIR(Batch Size 64)	**GPU**	**1.33s**	$\mathbf{4.4\text{x}10^{-5}}$**s**
FAST-RIR(Batch Size 128)	GPU	1.77s	$5.9\text{x}10^{-5}$s

Table 2. T_{60} error of our FAST-RIR for 30,000 testing acoustic environments. We report the T_{60} error for RIRs cropped at T_{60} and full RIRs. We only crop RIRs with T_{60} below 0.25s.

T_{60} Range	Crop RIR at T_{60}	T_{60} Error
0.2s - 0.25s	No	0.068s
0.2s - 0.25s	**Yes**	**0.033s**
0.25s - 0.7s	-	0.021s
0.2s - 0.7s	No	0.029s
0.2s - 0.7s	**Yes**	**0.023s**

Table 3. Far-field ASR results were obtained for far-field speech data recorded by single distance microphones (SDM) in the AMI corpus. The best results are shown in **bold**.

Training Dataset	Word Error Rate [%]	
Clean Speech ⊛ RIR	dev	eval
IHM ⊛ None	55.0	64.2
IHM ⊛ Image Method [5]	51.7	56.1
IHM ⊛ gpuRIR [13]	52.2	55.5
IHM ⊛ DAS [7]	47.9	**52.5**
IHM ⊛ DAS-cropped [7]	48.3	**52.6**
IHM ⊛ FAST-RIR (ours)	47.8	**53.0**

Discussion and Future Work

- We propose a novel FAST-RIR architecture to generate a large RIR dataset on the fly.
- We can easily train our FAST-RIR with the RIRs generated using any state-of-the-art RIR generator to improve the accuracy of RIR generation.
- We would like to expand our FAST-RIR to generate RIRs for any complex 3D scenes.