TS-RIR: TRANSLATED SYNTHETIC ROOM IMPULSE RESPONSES FOR SPEECH AUGMENTATION

Anton Ratnarajah, Zhenyu Tang, Dinesh Manocha

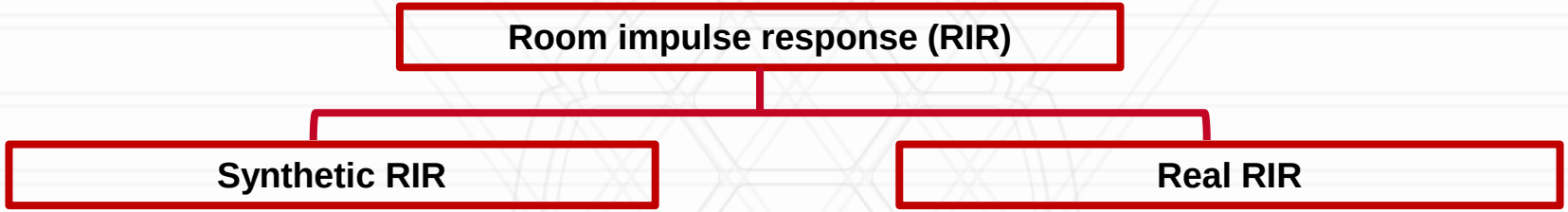
UNIVERSITY OF MARYLAND

Introduction

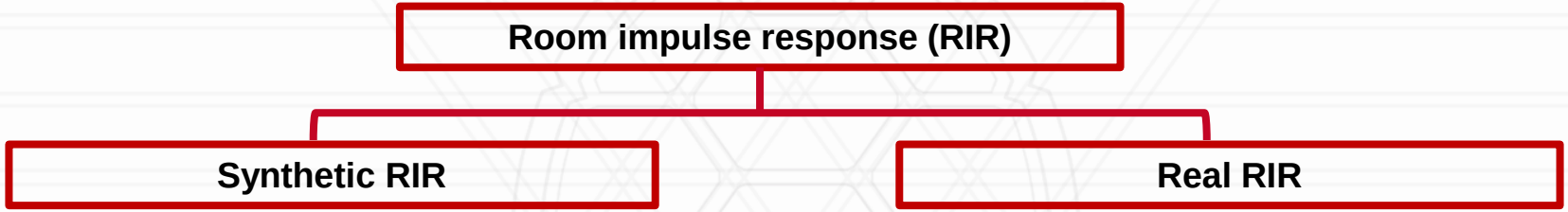
Far-field Speech = Reverberation Effects * Anechoic Speech + Background Noise

Room Impulse Response (RIR)

Introduction

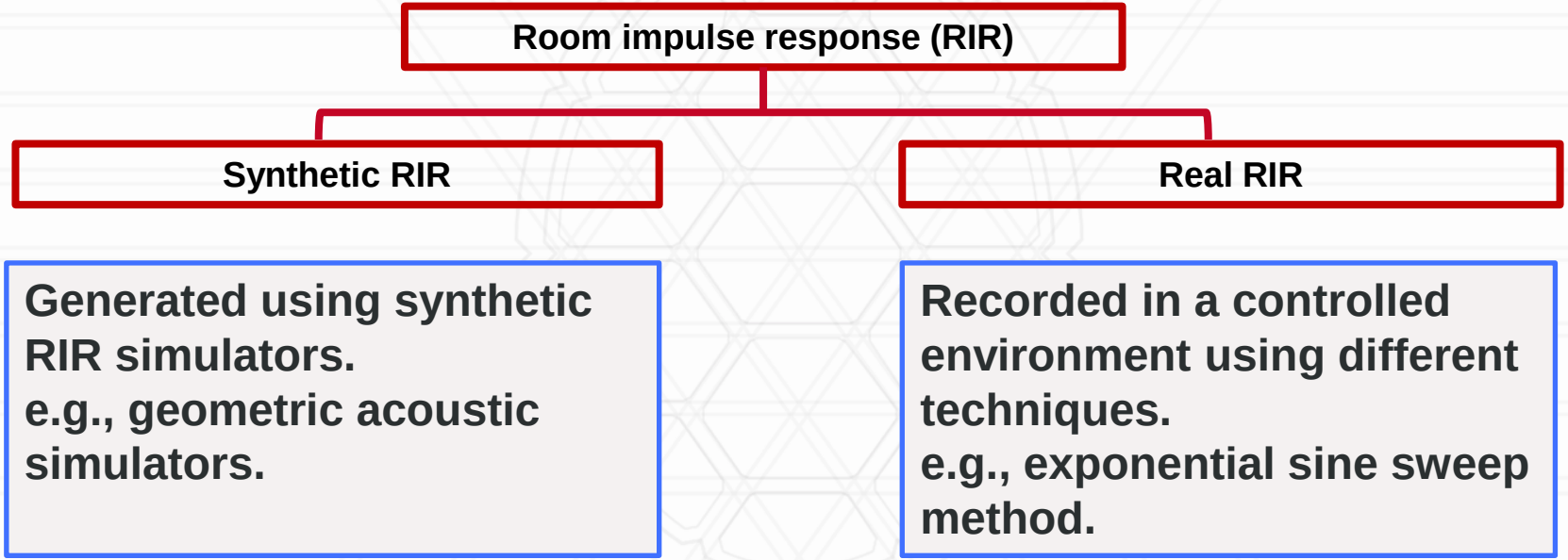


Introduction

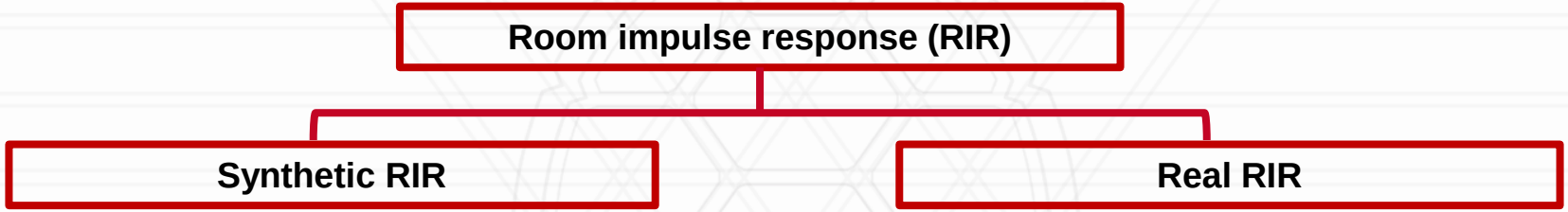


Generated using synthetic RIR generators. e.g., geometric acoustic simulators.

Introduction



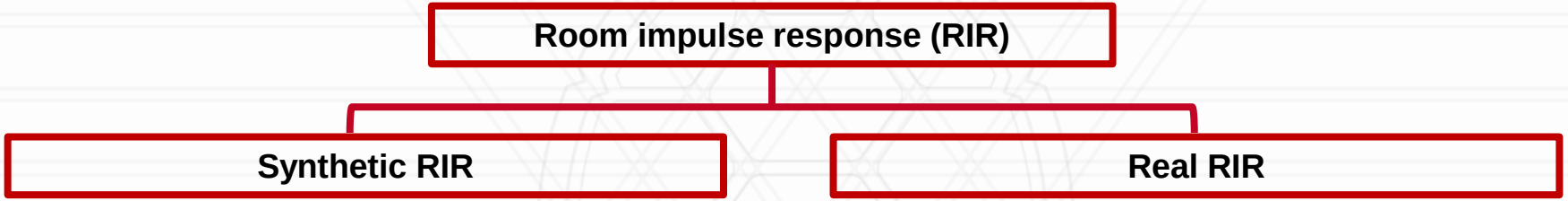
Advantage



Can generate infinite number of synthetic RIRs covering all possible scenarios.

Recorded RIRs can capture most of the details in the acoustic environment accurately (e.g., room materials).

Disadvantage



Cannot model all the reverberation effects accurately. For example, geometric acoustic simulators cannot model low-frequency wave effects such a diffractions and room resonance.

We need lot of human labor and special hardware to record real RIRs.

Proposed Method

- We propose an optimal post-processing approach to improve the quality of synthetic RIRs using real RIRs.

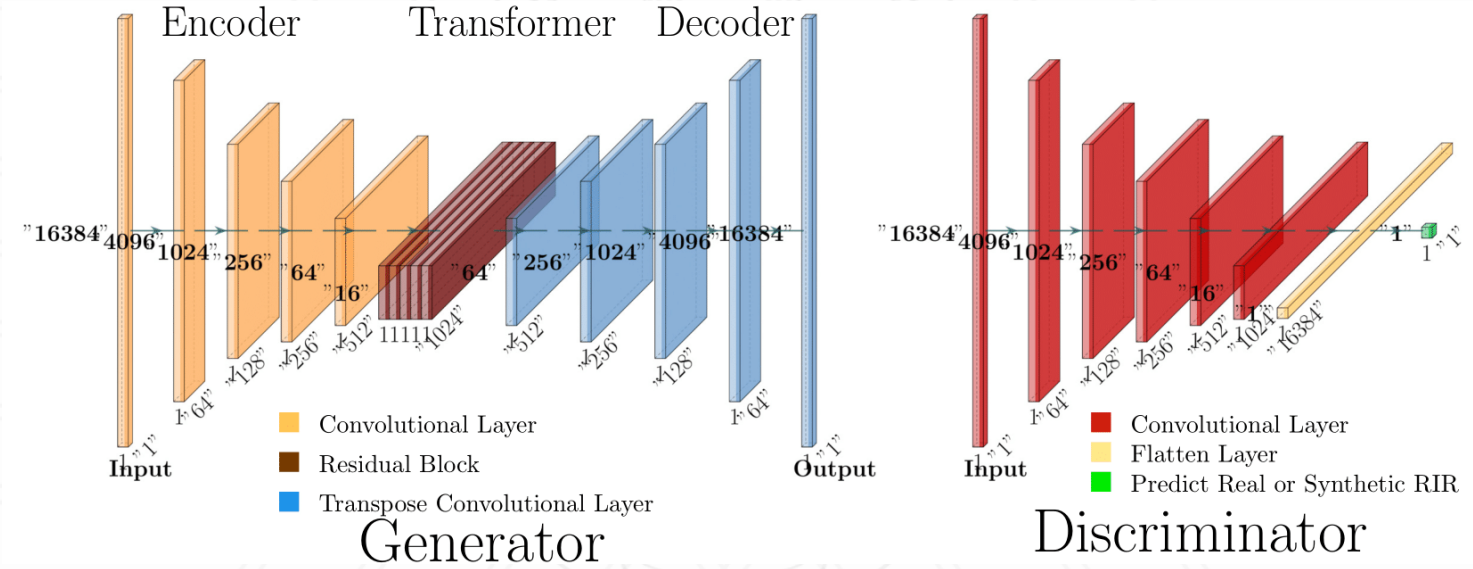
- We improve the quality of synthetic RIRs by compensating low frequency effects, similar to those in real RIRs.

Proposed Method

1. We translate synthetic RIRs to real RIRs using our proposed one-dimensional CycleGAN (TS-RIRGAN).

2. We perform real-world sub-band room equalization on the translated RIRs.

Translating synthetic RIRs to Real RIRs



Translating synthetic RIRs to Real RIRs

- We design a TS-RIRGAN architecture that learns mapping functions between one-dimensional synthetic RIRs and real RIRs in the absence of paired training examples.
- Inspired by WaveGAN, which applies generative adversarial networks (GANs) to raw-waveform audio, we directly input RIRs as raw audio samples to our network to learn the mapping functions.
- In most cases, real and synthetic RIRs are less than one second in duration. Therefore, we re-sample the synthetic and real RIR datasets without loss of generality to 16 kHz and pass them as a one-dimensional input of length 16384.
- Our objective function consists of adversarial loss, cycle-consistency loss and identity loss to learn the mapping functions.

Sub-band Room Equalization

- Sub-band room equalization bridges the gap in the frequency gain of real and synthetic RIRs over the entire frequency range.

- Our formulation is based on the sub-band room equalization approach described in [1]

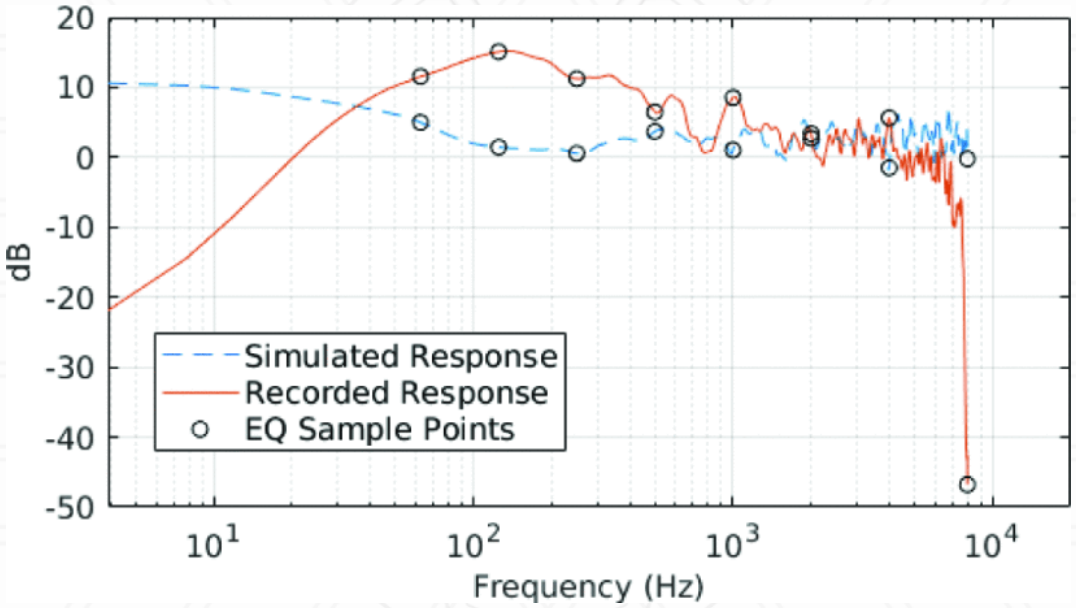
- Sub-band room equalization consists of 2 stages

    1. Sub-band relative gain calculation

    2. Equalization matching

[1]] Z. Tang, H. Meng, and D. Manocha, “Low-frequency compensated synthetic impulse responses for improved far-field speech recognition,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6974–6978.

Sub-band Room Equalization



- In this figure, we can observe that the frequency gains of the simulated response is relatively flat compared with those of the recorded response.

Sub-band relative gain calculation

- We compute the relative gain from the frequency response by taking the gain at 1000Hz as the reference for each real RIR in a real-world RIR dataset (BUT ReverbDB).
- Then we extract the relative frequency gain at 7 unique sample points (62.5Hz, 125Hz, 250Hz, 500Hz, 2000Hz, 4000Hz, 8000Hz) for every real RIR.
- The mean and standard deviations of the relative gains for each sample point are different. Therefore, we use a Gaussian mixture model to model 7 Gaussian distributions using the relative gains from the sampled points.
- We re-sample equal numbers of relative gains for each sample point as the input to the Gaussian mixture model.
- Instead of using the relative gains of the real RIRs, we use the re-sampled relative gains to avoid duplicating the real RIRs during equalization matching.

Equalization matching

- We compute the relative frequency gains for the **synthetic RIRs** at the chosen sample points (62.5Hz, 125Hz, 250Hz, 500Hz, 2000Hz, 4000Hz, 8000Hz), taking gain at 1000Hz as the reference.

- We calculate the difference in the relative gains of synthetic RIRs and the re-sampled relative gains.

- Next, we design a finite impulse response (FIR) filter using the window method [28] to compensate for the difference in the relative gains.

- We filter the synthetic RIRs using our designed FIR filter to match the sub-band relative gains of synthetic RIRs with the re-sampled relative gains.

Combination

- We tried different combinations of our post-processing to come up with the optimal combination.

Combination	Description
GAS+EQ	Only perform sub-band room equalization.
G_{SR}(GAS+EQ)	First, perform sub-band room equalization, then translate the equalized synthetic RIRs to real RIRs using our TS-RIRGAN.
G_{SR}(GAS)	Only translate synthetic RIRs to real RIRs using our TS-RIRGAN.
G_{SR}(GAS)+EQ	First, translate synthetic RIRs to real RIRs using our TS-RIRGAN, then perform sub-band room equalization to the translated RIRs.

Optimal Combination

- We choose the optimal combination of our post-processing approach based on

1. **Acoustic parameter values of the post-processed room impulse responses.**

2. **Energy distribution of the post-processed room impulse responses.**

Acoustic Parameters

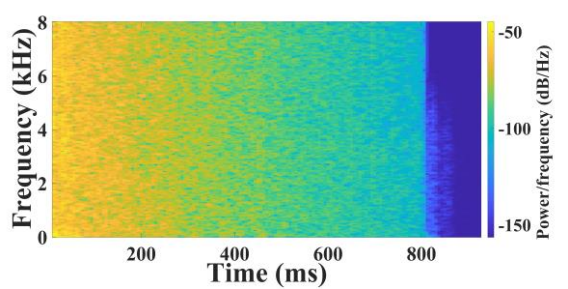
- Reverberation Time (T_{60})
- Direct to reverberant ratio (DRR)
- Early decay time (EDT)
- Early-to-late index (CTE)

Acoustic Parameters

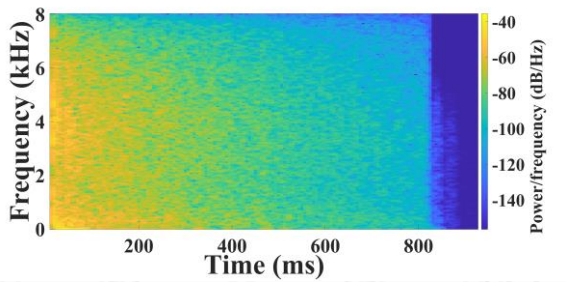
Table 1: Mean values of the acoustic parameters. We calculated the mean reverberation time (T60), mean direct-to-reverberant ratio (DRR), mean early-decay-time (EDT), and mean early-to-late index (CTE) for real, synthetic and post-processed synthetic RIRs. We also report the absolute mean difference of the acoustic parameters between synthetic and post-processed synthetic RIRs and real RIRs.

RIRs	T60 (seconds)		DRR (dB)		EDT (seconds)		CTE (db)	
	Mean	Diff	Mean	Diff	Mean	Diff	Mean	Diff
Real RIRs	1.0207		-6.3945		0.8572		3.4886	
GAS	**0.9553**	**0.0654**	-4.7277	1.6668	0.8846	0.0274	4.7536	1.265
GAS+EQ	0.9540	0.0667	-7.4246	1.0301	0.8912	0.0340	5.6404	2.1518
G_{SR}(GAS+EQ)	1.5493	0.5286	-8.3879	1.9934	1.046	0.1888	2.6562	0.8324
G_{SR}(GAS)	1.6433	0.6226	**-6.6491**	**0.2546**	**0.8483**	**0.0089**	**3.4907**	**0.0021**
G_{SR}(GAS)+EQ	1.6364	0.6157	**-6.7234**	**0.3289**	**0.8323**	**0.0249**	**3.5367**	**0.0481**

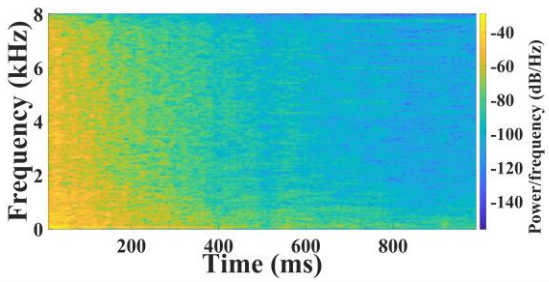
Spectogram



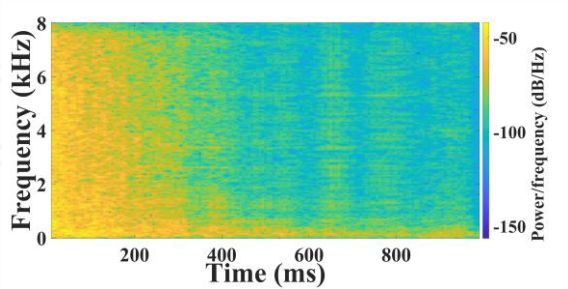
Synthetic RIR (GAS)



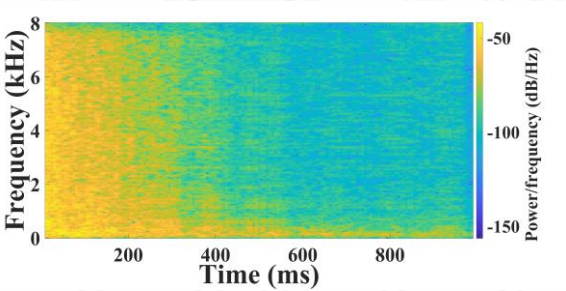
GAS+EQ



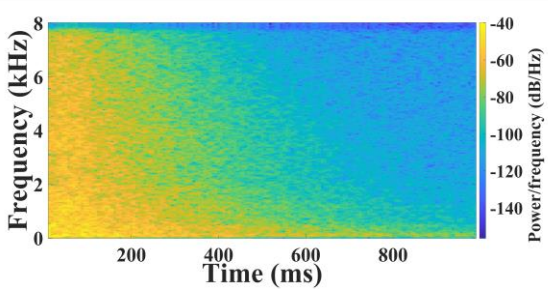
G_{SR}(GAS+EQ)



G_{SR}(GAS)



G_{SR}(GAS)+EQ



Real RIR (BUT ReverbDB)

ASR Experiment

- We evaluate our post-processed Synthetic RIRs on the Kaldi LibriSpeech far-field ASR recipe[1].
- We augment the far-field speech training set by convolving clean speech $\boldsymbol{x_c[t]}$ from LibriSpeech dataset with different sets of RIRs $\boldsymbol{r[t]}$ and adding environmental noise $\mathbf{n[t]}$ from BUT ReverbDB dataset.
- The environmental noise is started at a random position $\mathbf{l}$ and repeated in a loop to fill the clean speech.

$$\boldsymbol{x_f[t] = x_c[t] \circledast r[t] + \lambda * n[t + l]}$$

1) https://github.com/RoyJames/kaldi

ASR Experiment

- We train time-delay neural network on the augmented far-field speech training dataset.
- We extract the identity vectors (i-vectors) of the real-world far-field test set and decode using following language models.
  - Large four-gram (fglarge)
  - Large tri-gram (tglarge)
  - Medium tri-gram (tgmed)
  - Small tri-gram (tgsmall)
- We also do online decoding using tgsmall model. In online decoding, extracted features are passed in real-time instead of waiting until the entire audio is captured.

Experiments and Results

Table 1: Word error rate (WER) reported by the Kaldi far-field ASR system. We trained the Kaldi model using the different augmented far-field speech training sets and tested it on a real-world far-field speech. The training sets are augmented using synthetic RIRs (GAS), post-processed synthetic RIRs and real RIRs. We report WER for fglarge, tglarge, tgmed, and tgsmall phone language models and online decoding using tgsmall phone language model. Our best results are shown in **bold**.

Training Data		Test Word Error Rate (WER) [%]				
		fglarge	tglarge	tgmed	tgsmall	online
Clean (Baseline)		77.15	77.37	78.00	78.94	79.00
Real (Oracle)		12.40	13.19	15.62	16.92	16.88
GAS[1]		16.53	17.26	20.24	21.91	21.83
GAS+EQ[2]		14.51	15.37	18.33	20.01	19.99
Ours	**G_{SR}(GAS+EQ)**	14.27	14.98	17.79	19.37	19.36
	G_{SR}(GAS)	14.12	14.70	17.44	19.08	19.06
	G_{SR}(GAS)+EQ	**13.24**	**14.04**	**16.65**	**18.40**	**18.39**

[1] Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha, "Improving reverberant speech training using diffuse acoustic simulation," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6969– 6973.
[2]] Z. Tang, H. Meng, and D. Manocha, "Low-frequency compensated synthetic impulse responses for improved far-field speech recognition," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6974–6978.

Experiments and Results

Table 1: Word error rate (WER) reported by the Kaldi far-field ASR system. We trained the Kaldi model using the different augmented far-field speech training sets and tested it on a real-world far-field speech. The training sets are augmented using synthetic RIRs (GAS), post-processed synthetic RIRs, synthetic RIRs generated using IR-GAN and real RIRs. We report WER for fglarge, tglarge, tgmed, and tgsmall phone language models and online decoding using tgsmall phone language model. Our best results are shown in **bold**.

Training Data	Test Word Error Rate (WER) [%]				
	fglarge	tglarge	tgmed	tgsmall	online
Clean (Baseline)	77.15	77.37	78.00	78.94	79.00
Real (Oracle)	12.40	13.19	15.62	16.92	16.88
GAS[1]	16.53	17.26	20.24	21.91	21.83
GAS+EQ[2]	14.51	15.37	18.33	20.01	19.99
IR-GAN [3]	14.99	15.93	18.81	20.28	20.24
GAS+IR-GAN[3]	14.16	14.99	17.56	19.21	19.21
G_{SR}(GAS)+EQ (Ours)	**13.24**	**14.04**	**16.65**	**18.40**	**18.39**

[1] Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha, “Improving reverberant speech training using diffuse acoustic simulation,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6969– 6973.
[2]] Z. Tang, H. Meng, and D. Manocha, “Low-frequency compensated synthetic impulse responses for improved far-field speech recognition,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6974–6978.
[3Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha, “IR-GAN: Room Impulse Response Generator for FarField Speech Recognition,” in Proc. Interspeech 2021, 2021, pp. 286–290.

Summary

- We present a new architecture to translate synthetic RIRs to real RIRs and perform real-world sub-band room equalization on the translated RIRs to improve the quality of synthetic RIRs.

- We evaluate this post-processing approach on the Kaldi LibriSpeech far-field automatic speech recognition benchmark and observe that our proposed approach outperforms unmodified synthetic RIRs by up to **19.9%**.

UNIVERSITY OF
MARYLAND