IR-GAN: Room impulse generator for far-field speech recognition

Anton Ratnarajah, Zhenyu Tang, Dinesh Manocha

UNIVERSITY OF MARYLAND

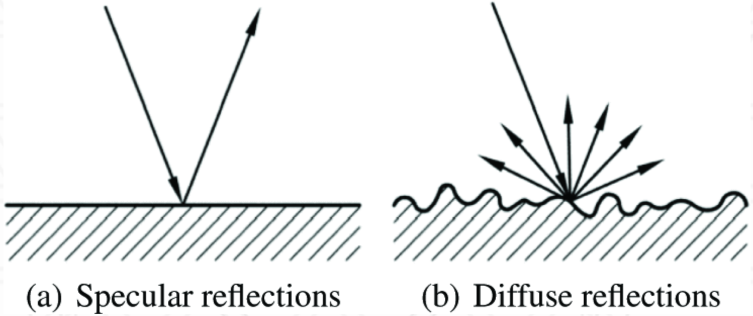
Introduction

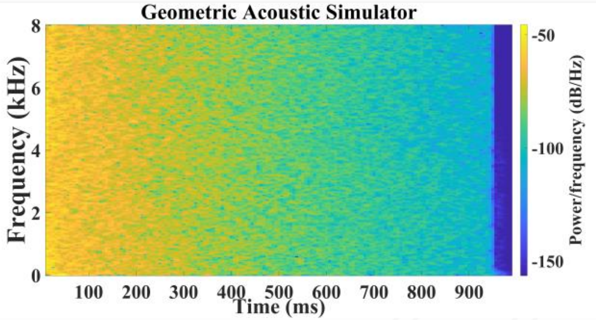
Clean Speech * Room Impulse Response + Environmental Noise = Far-field Speech

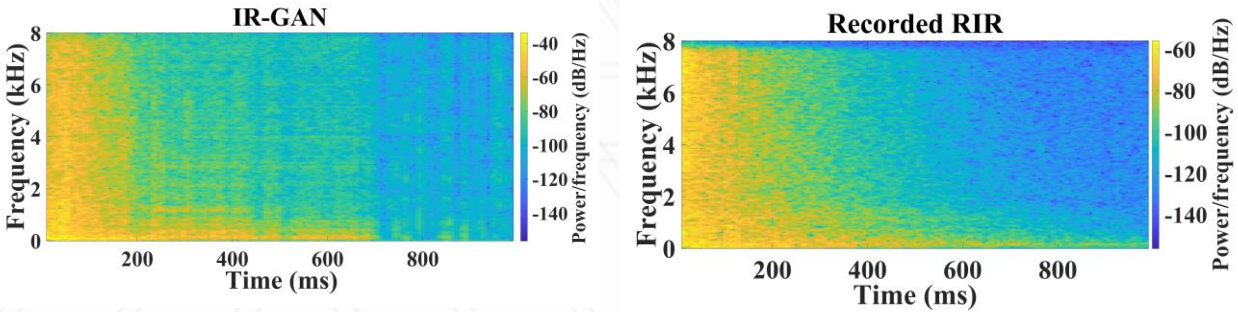
Related Works

- Room impulse response (RIR) generator
  - Physically-based acoustic simulator
    - Wave-based techniques
    - Geometric methods
      - Image method
      - Diffuse acoustic simulators
  - Neural network-based RIR generator
    - IR-GAN

Room Impulse Response







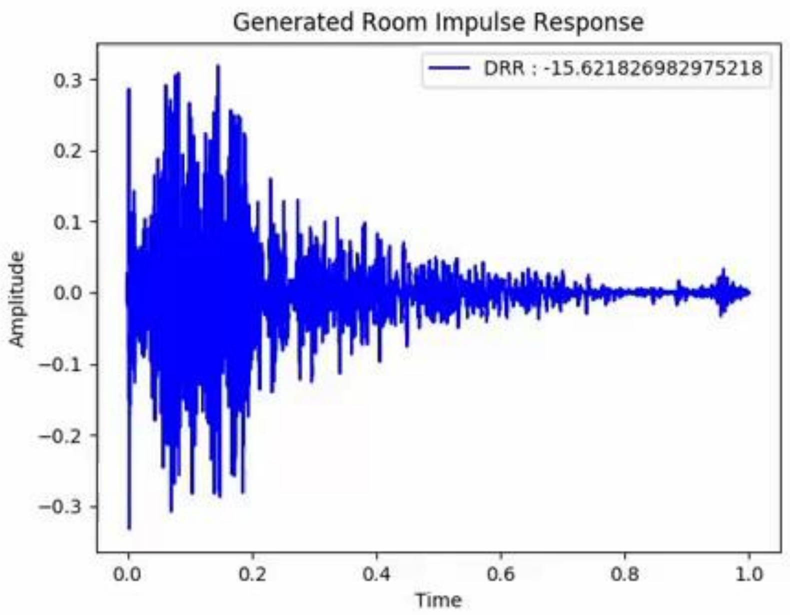
1) Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha, "Improving reverberant speech training using diffuse acoustic simulation," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6969– 6973.
2) I. Szoke, M. Sk ¨ acel, L. Mo ´ sner, J. Paliesek, and J. ˇ Cernock ˇ y,´ "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.

IR-GAN

- IR-GAN is a GAN-based room impulse response generator (RIR) that is trained on real-world RIRs.

- IR-GAN can generate RIRs corresponding to different acoustic environments by parametrically controlling following acoustic parameter
  1. Reverberation time (T_{60})
  2. Direct-to-reverberant ratio (DRR)
  3. Early-decay-time (EDT)
  4. Early-to-late index

IR-GAN



Generated Room Impulse Response
DRR : -15.621826982975218
Amplitude
Time
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.0
0.2
0.4
0.6
0.8
1.0

OUR APPROACH – Room Impulse Response (RIR) Representation

- **Input data**
  - Representation → Audio samples as a 32-bit floating-point vector
  - Sampling Rate → 16 kHz
  - Length → 16384 samples (slightly more than one second)

- **Output data**
  - Representation → Audio samples as a 32-bit floating-point vector
  - Sampling Rate → 16 kHz
  - Length → 16384 samples (slightly more than one second)

OUR APPROACH – Architecture

- **We adapt the WaveGAN [1] architecture to learn a mapping from low-dimensional vector space to a high-dimensional space where RIRs is represented.**

- **WaveGAN is a one-dimensional version of DCGAN [2] architecture where two-dimensional filters are replaced by one-dimensional filters.**

1) C. Donahue, J. McAuley, and M. Puckette, “Adversarial audio synthesis,” in ICLR, 2019.
2) A. R, L. M, and S. C, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, Y. B and Y. L, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.06434

OUR APPROACH – Value function

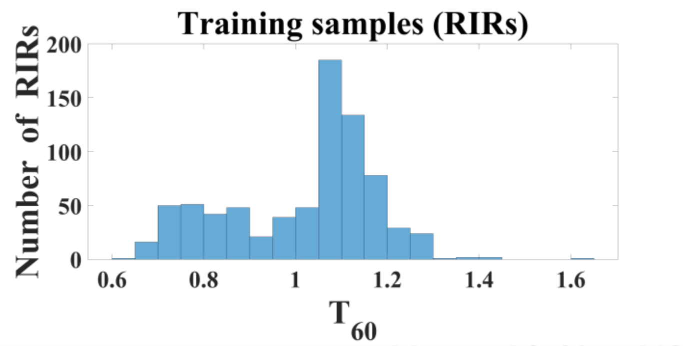
- As proposed in [1], we minimize the Wasserstein-1 distance between data distribution $\boldsymbol{p}_{\boldsymbol{data}(\boldsymbol{x})}$ and model distribution.

- Model distribution is implicit in the second part of the equation because $\boldsymbol{G}(\boldsymbol{z})$ represents the mapping from a latent vector $\boldsymbol{z}$ with distribution $\boldsymbol{p}_{\boldsymbol{z}(\boldsymbol{z})}$ to the data space.

$$\boldsymbol{V}_{\boldsymbol{WGAN}}(\boldsymbol{D}_{\boldsymbol{WGAN}}, \boldsymbol{G}) = \boldsymbol{E}_{\boldsymbol{x} \sim \boldsymbol{p}_{\boldsymbol{data}(\boldsymbol{x})}}[\log \boldsymbol{D}_{\boldsymbol{WGAN}}(\boldsymbol{x})] - \boldsymbol{E}_{\boldsymbol{z} \sim \boldsymbol{p}_{\boldsymbol{z}(\boldsymbol{z})}}[\boldsymbol{log}\, \boldsymbol{D}_{\boldsymbol{WGAN}}(\boldsymbol{G}(\boldsymbol{z}))].$$

1) M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 214–223. [Online]. Available: http://proceedings.mlr.press/v70/arjovsky17a.html.

OUR APPROACH – Constrained Room Impulse Response generation



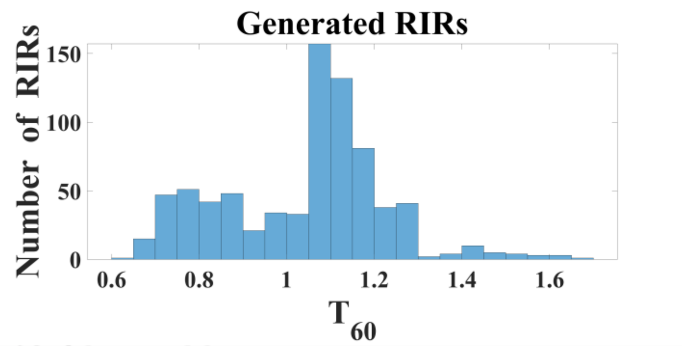
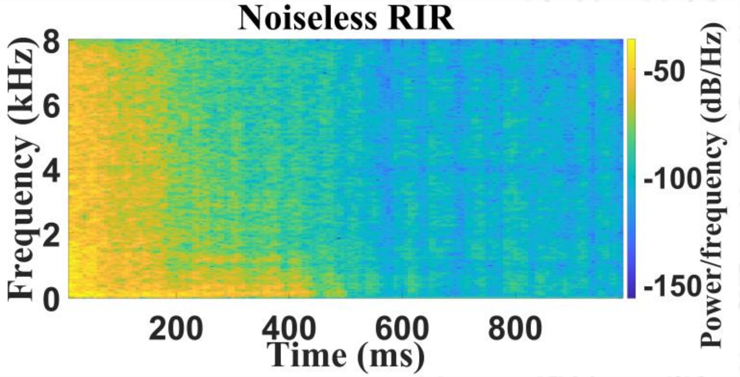


Figure: T_{60} distribution of training samples and T_{60} distribution of RIRs generated using our IR-GAN with the constraint.

OUR APPROACH – Constrained Room Impulse Response generation



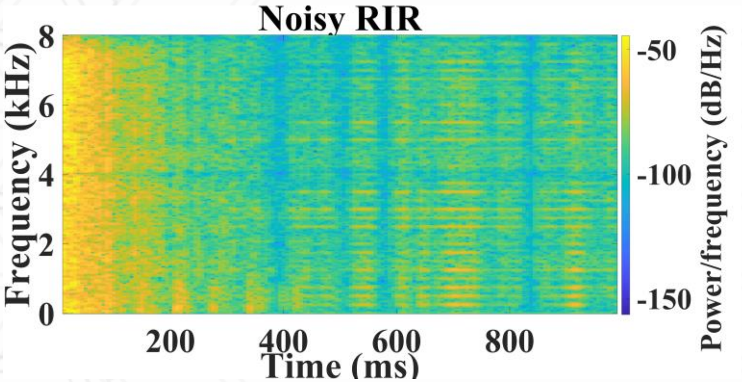


Figure: Spectrogram of noiseless RIR and noisy RIR. The noiseless RIR has a T_{60} value of around 1, and the noisy RIR has a T_{60} value of around 3. In the noisy spectrogram, we can see many horizontal artifacts around 700ms.

ASR Experiment

- We evaluate our post-processed Synthetic RIRs on the Kaldi LibriSpeech far-field ASR recipe.
- We augment far-field speech training set by convolving clean speech $\boldsymbol{x_c}[\boldsymbol{t}]$ from LibriSpeech dataset with different sets of RIRs $\boldsymbol{r}[\boldsymbol{t}]$ and adding environmental noise $\mathbf{n}[\mathbf{t}]$ 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.

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

1) GitHub - RoyJames/kaldi: This is a Kaldi fork for modified LibriSpeech reverb training.

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 go online decoding using tgsmall model. In online decoding, extracted features are passed in real-time instead of waiting until the entire audio is captured.

1) GitHub - RoyJames/kaldi: This is a Kaldi fork for modified LibriSpeech reverb training.

Experiments and Results

Table 1: Different RIRs used in our experiment.

RIR	Description
BUT	Real-world RIRs from the BUT ReverbDB dataset [1]
AIR	Real-world RIRs from the AIR [2] dataset.
GAS	Simulated RIRs using the acoustic simulator [3].
GAN.C	RIRs generated using our IR-GAN with constraint
GAN.U	RIRs generated using our IR-GAN without any constraint.

1) I. Szoke, M. Sk ¨ acel, L. Mo ´ sner, J. Paliesek, and J. ˇ Cernock ˇ y,´ “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.
2) 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, 2009, pp. 1–5.
3) 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.

Experiments and Results

Dataset	RIR	Hours	Number of RIRs	Librispeech Dataset	Noise
Test Dataset	BUT	5.4	242	test-clean	BUT
	AIR	5.4	68	test-clean	BUT
Training Dataset	BUT	460	773	train-clean-{100,360)	BUT
	GAS	460	773	train-clean-{100,360)	BUT
	GAN.C	460	773	train-clean-{100,360)	BUT
	GAN.U	460	773	train-clean-{100,360)	BUT
	GAN.C + GAS	460	1546	train-clean-{100,360)	BUT
	2*GAN.C	460	1546	train-clean-{100,360)	BUT

Experiments and Results

Table 3: Far-field automatic speech recognition results obtained from the far-field LibriSpeech test set. In this table, *BUT and *AIR represent far-field test sets generated using real RIRs from the BUT ReverbDB and AIR datasets, respectively. clean* represents clean speech. WER is reported for the tri-gram phone (tglarge, tgmed, tgsmall) and four-gram phone (fglarge) language models, and online decoding using tgsmall. Best results in each comparison are marked in **bold**.

Experimental Setup (training set) @ (test set)	Test Word Error Rate (WER) [%]				
	fglarge	tglarge	tgmed	tgsmall	online
Clean @ BUT (Baseline)	77.15	77.37	78.00	78.94	79.00
BUT @ BUT (Oracle) [1]	12.40	13.19	15.62	16.92	16.88
GAS @ BUT [2]	16.53	17.26	20.24	21.91	21.83
GAN.U @ BUT	19.71	20.74	24.27	25.93	25.90
GAN.C @ BUT	**14.99**	**15.93**	**18.81**	**20.28**	**20.24**

1) I. Szoke, M. Sk ̈acel, L. Mo ́sner, J. Paliesek, and J. ̌Cernock ̌y, ́ "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.
2) 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.
3) 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, 2009, pp. 1–5.

Experiments and Results

Table 3: Far-field automatic speech recognition results obtained from the far-field LibriSpeech test set. In this table, *BUT and *AIR represent far-field test sets generated using real RIRs from the BUT ReverbDB and AIR datasets, respectively. clean* represents clean speech. WER is reported for the tri-gram phone (tglarge, tgmed, tgsmall) and four-gram phone (fglarge) language models, and online decoding using tgsmall. Best results in each comparison are marked in **bold**.

Experimental Setup (training set) @ (test set)	Test Word Error Rate (WER) [%]				
	fglarge	tglarge	tgmed	tgsmall	online
Clean @ BUT (Baseline)	77.15	77.37	78.00	78.94	79.00
BUT @ BUT (Oracle) [1]	12.40	13.19	15.62	16.92	16.88
GAN.C @ BUT	14.99	15.93	18.81	20.28	20.24
2*GAN.C @ BUT	14.86	15.69	18.50	20.25	20.17
GAN.C+GAS @ BUT	**14.16**	**14.99**	**17.56**	**19.21**	**19.21**

1) I. Szoke, M. Sk ¨acel, L. Mo ´ sner, J. Paliesek, and J. ˇ Cernock ˇ y,´ "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.
2) 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.
3) 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, 2009, pp. 1–5.

Experiments and Results

Table 3: Far-field automatic speech recognition results obtained from the far-field LibriSpeech test set. In this table, *BUT and *AIR represent far-field test sets generated using real RIRs from the BUT ReverbDB and AIR datasets, respectively. clean* represents clean speech. WER is reported for the tri-gram phone (tglarge, tgmed, tgsmall) and four-gram phone (fglarge) language models, and online decoding using tgsmall. Best results in each comparison are marked in **bold**.

Experimental Setup (training set) @ (test set)	Test Word Error Rate (WER) [%]				
	fglarge	tglarge	tgmed	tgsmall	online
Clean @ AIR	26.79	27.40	29.64	30.88	31.15
GAN.C @ AIR	**7.71**	**8.03**	**9.88**	**11.11**	**11.08**

1) I. Szoke, M. Sk ¨ acel, L. Mo ´ sner, J. Paliesek, and J. ˇ Cernock ˇ y,´ "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.
2) 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.
3) 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, 2009, pp. 1–5.

Summary

- We present GAN based Room Impulse Response (RIR) generator to generate realistic RIRs using acoustic parameters.

- Our IR-GAN outperforms the state-of-the-art geometric acoustic simulator by upto **8.95%**.

- We show that combining synthetic data generated using IR-GAN with existing geometric acoustic simulator can boost the performance of the far-field ASR system.

UNIVERSITY OF
MARYLAND