INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

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

Anton Ratnarajah, Zhenyu Tang, Dinesh Manocha

University of Maryland, College Park, MD 20742, USA

jeran@umd.edu,zhy@umd.edu,dmanocha@umd.edu

Abstract

We present a Generative Adversarial Network (GAN) based room impulse response generator (IR-GAN) for generating realistic synthetic room impulse responses (RIRs). IR-GAN extracts acoustic parameters from captured real-world RIRs and uses these parameters to generate new synthetic RIRs. We use these generated synthetic RIRs to improve far-field automatic speech recognition in new environments that are different from the ones used in training datasets. In particular, we augment the far-field speech training set by convolving our synthesized RIRs with a clean LibriSpeech dataset [1]. We evaluate the quality of our synthetic RIRs on the far-field LibriSpeech test set created using real-world RIRs from the BUT ReverbDB [2] and AIR [3] datasets. Our IR-GAN reports up to an 8.95% lower error rate than Geometric Acoustic Simulator (GAS) in far-field speech recognition benchmarks. We further improve the performance when we combine our synthetic RIRs with synthetic impulse responses generated using GAS. This combination can reduce the word error rate by up to 14.3% in far-field speech recognition benchmarks.

Index Terms: acoustic simulation, room impulse response, generative adversarial network, speech recognition

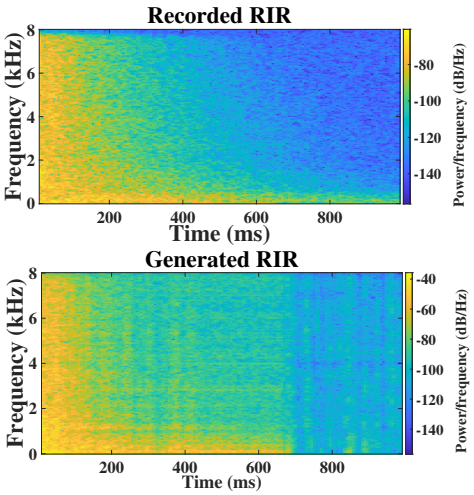

Figure 1: *Spectrogram of real RIR and RIR generated using our GAN-based approach. We can see both spectrograms have similar energy distributions.*

1. Introduction

Reverberation is a part of the speech signal which characterizes the acoustic environment (e.g., room geometry, loudspeaker and microphone location, room materials, etc.) used to capture the speech signal. Reverberation can be characterized by the transfer function known as the room impulse response (RIR). A room impulse response represents the relationship between the dry sound and the reflection of the sound signal from the boundaries of the room [4].

RIRs are frequently used in many practical applications such as far-field speech recognition [2, 5, 6], speech enhancement [7], speech separation [8], sound rendering [9], audio forensics [10], etc. One challenge for these applications is that existing recorded or real-world RIR datasets are collected in limited acoustic environments. In our paper, we address this issue by augmenting RIRs covering a wide range of acoustic environments using a Generative Adversarial Network (GAN).

Many prior techniques for generating synthetic RIRs are based on acoustic simulators [11, 12]. These simulators use room geometry, sound absorption, and sound reflection coefficients as input and generate RIRs by simulating occlusion, specular reflections and diffuse reflections. In practice, such synthetic RIR generators can model some sound propagation phenomena in regularly shaped or mostly empty rooms. On the other hand, simulating the sound reverberation effects in complex scenarios like stairways is more difficult [5]. As a result, we need other synthetic RIR generator methods that can model the sound effects in complex environments.

Main Contributions: We present a novel GAN-based RIR generator (IR-GAN) that is trained on real-world RIRs. IR-GAN can parametrically control different acoustic parameters (e.g., reverberation time, direct-to-reverberant ratio, etc.) learned from real RIRs and generate synthetic RIRs that can imitate new or different acoustic environments. Moreover, we propose a constrained RIR generation approach that can avoid synthesizing RIRs with noisy artifacts to a greater extent.

Our IR-GAN uses WaveGAN [13] architecture to synthesize new RIRs by learning from real RIRs. IR-GAN maps all acoustic parameters in real RIRs to a high-dimensional space and generates a wide range of RIRs by controlling various acoustic parameters. As a result, we can train on real-world RIRs corresponding to complex locations like stairways and can augment RIRs corresponding to such locations. Figure 1 highlights the spectrogram of a real RIR from the BUT ReverbDB dataset [2] and the spectrogram of a generated RIR using our approach.

We create far-field speech to evaluate our RIRs in a far-field automatic speech recognition (ASR) system. Our far-field speech $x_f[t]$ is augmented by convolving clean speech $x[t]$ from the LibriSpeech dataset [1] with RIRs $h[t]$ and adding background noise $n[t]$ from the BUT ReverbDB [2] dataset. The starting position k of the noise is selected randomly, and the noise is repeated in a loop to fill clean speech. The weight α is calculated for a random signal-to-noise ratio within the range 10 to 100 for each far-field speech. We use real RIRs from the BUT ReverbDB [2] and AIR [3] datasets to create real-world far-field speech.

$$x_f[t] = x[t] \circledast h[t] + \alpha * n[t+k]. \quad (1)$$

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-230

Our GAN-based synthetic RIR generation approach is complementary to prior synthetic RIR generators based on acoustic simulators. We evaluate the performance by conducting far-field ASR tests and show that combining RIRs synthesized from IR-GAN and a state-of-the-art geometric acoustic simulator reduces the word error rate by up to 14.3%. Our code is published for follow-up research [1].

2. Related Works

Physically-based acoustic simulators have been used over the decades to generate synthetic RIRs for far-field speech research. Wave-based methods and geometric methods are widely used to model RIRs for different acoustic environments. The wave-based approach [14] solves the wave equation using numerical methods. Although wave-based methods accurately compute the RIRs, these methods are computationally expensive and only feasible for low frequencies and less complex scenes. The image source method [15] and path tracing methods [12, 16] are common geometric acoustic simulation-based methods. The image source method only models specular reflections in simple rectangular rooms while path tracing-based geometric acoustic simulators model occlusion and specular and diffuse reflections. Geometric acoustic simulators treat sound waves in the form of a ray [17]. Although this assumption holds for high-frequency waves, ray assumption causes visible irregularities at low frequencies [5]. In recent works, TS-RIRGAN [18] is used to transfer low-frequency wave effects learned from real RIRs to synthetic RIRs generated using geometric acoustic simulators. In many scenarios, the target room environment (i.e., the exact geometric shape and material parameters) is unknown or too complex for the geometric acoustic simulators. As a result, their ability to generate RIRs for all kinds of scenarios can be limited.

To overcome the limitations of synthetic RIRs, real RIRs are recorded in a controlled environment. The maximum length sequence method [19], the time-stretched pulses method [20], and the exponential sine sweep method [21] are common methods to measure real RIRs. Among these approaches, the exponential sine sweep method is robust to changing loudspeaker output volume and performs well in automatic speech recognition tasks. The real RIRs in BUT ReverbDB [2] are collected using the exponential sine sweep method. Since collecting real RIRs is time-consuming and technically difficult, only a limited number of real RIRs is available to augment far-field speech.

GANs have made steady progress over the years in image generation [22], image inpainting [23], and domain adaptation [24]. The success of GAN in computer vision motivated researchers to use it in other fields. Recently, GANs have been becoming popular in the audio generation. GANs have shown progress from music generation [25] to any short audio clip generation [13, 26]. In this work, we aim to augment high-quality RIRs using existing real RIRs. We use GANs for RIR generation to complement the prior works.

3. Our Proposed Approach: IR-GAN

3.1. Room Impulse Response Statistics

Room impulse response acoustic parameters are used to characterize the acoustic environment [27] and control RIR generation using GAN. Reverberation time (T_{60}), direct-to-reverberant ratio (DRR), early-decay-time (EDT), and early-to-late index (CTE) are four acoustic parameters that can be estimated from RIRs. We use these acoustic parameters to constrain IR-GAN-based RIR augmentation. Reverberation time measures the amount of time taken to decay the sound pressure by 60 decibels (dB). The T_{60} value depends on room size and the characteristics of the material (e.g., floor, walls, furniture, etc.). DRR is calculated by dividing the sound pressure level of a direct sound source by the sound pressure level of the sound arriving after one or more surface reflections [28]. DRR is measured in dB. Time taken for sound pressure to decay by 10 dB is multiplied by a factor of 6 to get early-decay-time. EDT depends on the type and location of the sound source. CTE measures the proportion of the total sound energy received in the first $50ms$ to the energy received during the rest of the period [4].

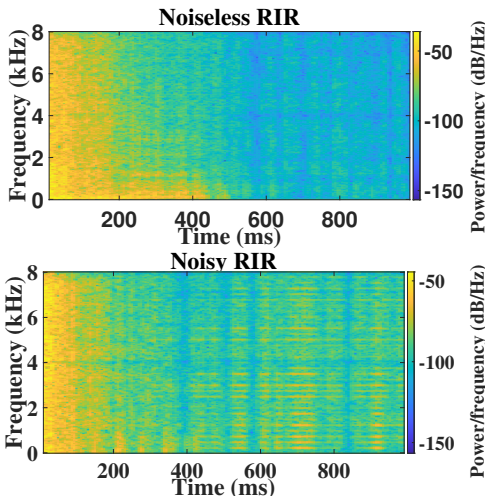

Figure 2: *Spectrogram of noiseless RIR and noisy RIR. The noiseless RIR has a T_{60} value of around 1, and the noisy RIR has a T_{60} value of around 3. In the noisy spectrogram, we can see many horizontal artifacts around 700ms.*

3.2. Room Impulse Response Representation

The representation of the input data and the data generated by the neural network is important for synthesizing high-quality RIRs. Therefore, lossy representations of RIRs like Mel-frequency cepstral coefficients (MFCCs) are less favorable. Audio samples are a lossless representation that can be easily converted to an audio signal. As different datasets store RIRs with different sampling rates, we re-sample all the RIRs to 16 kHz. We pass audio samples as a 32-bit floating-point vector of length 16384 to the GAN. The vector length is sufficient to represent RIRs because most of the real RIRs are less than one second in duration. We can represent slightly more than one second with 16384 samples with a sampling rate of 16 kHz.

3.3. GAN

GAN is a generative model that learns a mapping from a low-dimensional vector space to a high-dimensional space where the data is represented. We adapt the WaveGAN architecture proposed in [13] to generate high-quality RIRs. WaveGAN is a one-dimensional version of DCGAN [29] where two-dimensional filters are replaced by one-dimensional filters.

GANs trained using the value function proposed in the orig-

[1] `https://gamma.umd.edu/pro/speech/ir-gan` (with video)

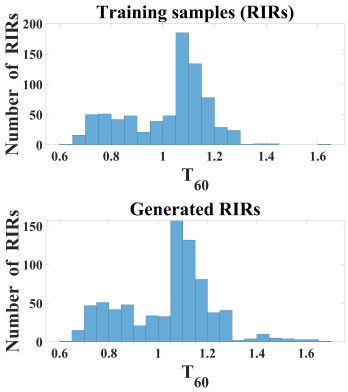

Figure 3: T_{60} *distribution of training samples and* T_{60} *distribution of RIRs generated using our IR-GAN with the constraint.*

inal GAN paper [30] are often unstable, and mode collapse can occur when the generator architecture is varied. Therefore, we use a stable cost function introduced in WGAN [31]. In this cost function, we minimize the Wasserstein-1 distance between data distribution p_{data}(x) and model distribution (Equation 2). Model distribution is implicit in the second part of the equation because G(z) represents the mapping from a latent vector z with distribution p_z(z) to the data space. In WGAN, the discriminator network D_{WGAN} gives a score based on the realness of the given image instead of predicting the probability that x comes from the real distribution. In Equation 2, E represents expectation.

$$V_{WGAN}(D_{WGAN}, G) = E_{x\sim p_{data}(x)}[\log D_{WGAN}(x)] - E_{z\sim p_z(z)}[\log D_{WGAN}(G(z))]. \quad (2)$$

3.4. Constrained RIR Generation

In our approach, we train a GAN to learn the mapping from the 100-dimensional latent vector z drawn from a Gaussian distribution to the RIR in data space. As the number of real-world RIR datasets is limited, we propose a constrained generation of RIRs from the generator network.

There is an infinite possibility to generate a 100-dimensional vector where each dimension can take any floating-point number between -1 and 1. Since we train GAN with a limited number of RIRs in real RIR datasets (BUT ReverbDB [2] contains less than 2000 RIRS.), there is a chance that some of the latent vectors map to noisy RIRs. For example, GAN may generate RIRs with unrealistically large T_{60} values. In Figure 2, we can see a noisy RIR generated without any constraint. The generated noisy RIR with a T_{60} value of around 3 has many horizontal artifacts around 700ms.

To prevent such mappings, we calculate the key acoustic parameters of the training samples (T_{60}, DRR, CTE, and EDT) and use them to generate histograms. Later, we generate RIRs by constraining them to fit the distribution of key acoustic parameters as the training samples. In this way, we can avoid noisy mapping to a greater extent. Since it is difficult to match the exact distributions of the training samples, we relax GAN to generate samples closer to the distribution with low probability. Figure 3 depicts the T_{60} distribution of the training samples and the T_{60} distribution of RIRs generated with the constraint.

Table 1: *Different RIRs used in our experiment.*

RIR	Description
BUT	Real-world RIRS from the BUT ReverbDB dataset [2].
AIR	Real-world RIRS from the AIR [3] dataset.
GAS	Simulated RIRs using the acoustic simulator [12].
GAN.C	RIRs generated using our IR-GAN with constraint (§ 3.4).
GAN.U	RIRs generated using our IR-GAN without any constraint.

Table 2: *Detailed information about the augmented dataset generated using different RIRs (Table 1). GAN.C+GAS indicates an equal mixture of GAS and GAN.C synthesized RIRs. 2*GAN.C contains twice the number of RIRs when compared to GAN.C.*

Dataset	RIR	Hours	#RIRs	LibriSpeech Dataset
Test Dataset	BUT	5.4	242	test-clean
	AIR	5.4	68	test-clean
Training Dataset	BUT	460	773	train-clean-{100,360}
	GAS	460	773	train-clean-{100,360}
	GAN.C	460	773	train-clean-{100,360}
	GAN.U	460	773	train-clean-{100,360}
	GAN.C+GAS	460	1546	train-clean-{100,360}
	2*GAN.C	460	1546	train-clean-{100,360}

4. Experiments and Results

We evaluate the effectiveness of our proposed approach by conducting far-field automatic speech recognition (ASR) experiments using the modified Kaldi LibriSpeech ASR recipe[2]. We use augmented far-field speech to train and test the Kaldi LibriSpeech ASR model and evaluate the benefits in the following manner. First, we compare the performance of our proposed IR-GAN with the state-of-the-art synthetic RIR generator [12]. Second, we evaluate the robustness of our IR-GAN when we train the GAN on one dataset [2] and test the GAN on another dataset [3] from different acoustic environments. We use word error rate to evaluate the performance.

4.1. Data Preparation

As proposed in [5], we generate far-field speech from clean LibriSpeech by convolving it with RIRs and adding environmental noise. Since we mainly focus on the quality of the synthesized RIRs, we use the same environmental noise from the BUT ReverbDB [2] for training and test set generation.

We use real RIRs from the BUT ReverbDB dataset [2] and the AIR [3] dataset to conduct our experiments. BUT ReverbDB consists of 1891 RIRs and 9114 environmental noises covering nine different rooms. To make fair comparisons, we use the 1209 BUT ReverbDB RIRs picked in [5]. The AIR dataset consists of 344 real RIRs from 6 different rooms. From these, we select 68 RIRs from the 4 rooms (studio booth, office room, lecture room, and meeting room) mentioned in [3]. We split 1209 real-world RIRs from the BUT ReverbDB dataset into subsets of {773,194,242} to generate training, development, and test far-field speech datasets. To test the robustness of our proposed approach, we use 68 RIRs from the AIR dataset.

Table 1 describes different RIRs used to create far-field

[2]https://github.com/RoyJames/kaldi.

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**.*

Experiment 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)	12.40	13.19	15.62	16.92	16.88
GAS@BUT [12]	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**
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**
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**

speech datasets in our experiment. Table 2 shows the detailed composition of the augmented datasets. The augmentation process does not change the overall duration of the original LibriSpeech dataset.

4.2. Synthetic RIR generation

For a fair comparison, we use the same synthetic RIRs generated using the state-of-the-art geometric acoustic simulator [12] as the previous benchmark. The geometric acoustic simulator synthesizes RIR using the meta-info provided in BUT ReverbDB. The meta-info includes room dimensions and loudspeaker and microphone locations. Therefore, the synthesized RIRs for the training and development sets mimic real RIR training and development sets to some extent.

We train our IR-GAN with 967 real RIRs from the BUT ReverbDB dataset. They are composed of real-world RIRs allocated for training and development. We generate synthetic RIRs with and without the constrained RIR generation process in § 3.4.

4.3. ASR Experiment

We use the modified Kaldi LibriSpeech ASR recipe to conduct our ASR experiments. We train time-delay neural networks [32] for each of our augmented far-field training sets. We extract the identity vectors [33] (i-vectors) of the real-world far-field test set and decode them using large four-gram (fglarge), large tri-gram (tglarge), medium tri-gram (tgmed), and small tri-gram (tgsmall) phone language models. We also do online decoding on the tgsmall phone language model. In online decoding, extracted features are passed in real-time instead of waiting until the entire audio is captured. Word error rate (WER) of each of the language model is used to evaluate the synthesized RIRs.

All the training and testing is done on 32 Intel(R) Xeon(R) Silver 4208 CPUs @ 2.10 GHz and 2 GeForce RTX 2080 Ti GPUs. For a fair comparison, we generate all the results from the same environment. It takes around four days to prepare the dataset and conduct each experiment on the Kaldi toolkit.

4.4. Results

Table 3 presents the ASR test WER for far-field speech generated using the BUT ReverbDB [2] and AIR [3] datasets. WER is calculated for four different phone language models (fglarge, tglarge, tgmed, and tgsmall) in Kaldi as well as for online decoding using a tgsmall phone language model.

We use WER to measure the robustness of the trained model. A lower WER indicates that the trained model shows superior accuracy in test conditions. Robustness depends on the model architecture and the input data used to train the model. In our experiments, we keep the model constant while we train with different datasets. Different training datasets are created by convolving the same LibriSpeech speech corpus with different RIRs. Therefore, the robustness of the model is only affected by the RIRs being used to generate the training datasets. As expected, a significantly high WER is reported when we train our baseline model on clean speech and test on the real RIRs. The lowest test WER is reported when we train and test on the real RIRs.

In Table 3, we can see that the proposed IR-GAN (GAN.C) gives a lower WER than the state-of-the-art geometric acoustic simulator (GAS). Lower WER indicates that the RIRs synthesized using IR-GAN are more realistically synthesized than the RIRs computed using the physically-based acoustic simulator. When we look at the WER for the fglarge model, we can see that our proposed IR-GAN gives an 8.95% lower error rate than the GAS. We can see that the RIRs generated using the unconstrained IR-GAN (GAN.U) performs poorly in our far-field speech recognition experiment. Therefore our constrained RIR generation approach is important to eliminate noisy RIR generation.

Hybrid Combination: Because the IR-GAN and GAS try to mimic real RIRs using two different approaches, we evaluate the WER when trained using a combination of synthetic RIRs generated using the IR-GAN and GAS. We observe that there is a further drop of up to 5% in WER compared to doubling the synthetic RIRs from the IR-GAN. The drop in WER indicates that we can boost the robustness of ASR systems by combining RIRs generated from our IR-GAN and GAS.

In practical scenarios, we do not have real RIRs from the acoustic environment where we need the capabilities for far-field ASR. Therefore, we augment RIRs using our IR-GAN trained with real RIRs from BUT ReverbDB [2]. Then we train the Kaldi LibriSpeech ASR model on far-field speech generated using the augmented RIRs and test this ASR model on the far-field speech augmented using the AIR dataset [3]. We can observe around a 19% absolute reduction in error when compared to training an ASR model with clean speech.

5. Discussion and Future Work

In this paper, we present an IR-GAN to generate realistic RIRs. Our proposed approach outperforms the state-of-the-art geometric acoustic simulator (GAS) by up to 8.95% in far-field ASR tests. When we combine our RIRs with RIRs generated using GAS, we can see a total reduction in word error rate by up to 14.3% in far-field ASR tests. This reduction in word error rate indicates that synthetic data generated using IR-GAN and GAS can be combined to boost the performance of far-field ASR systems. We have tested our approach only on indoor scenes, and extending it to outdoor scenes is a good topic for future work.

6. References

[1] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[2] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, "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.

[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.

[4] T. E. Vigran, *Building Acoustics*. CRC Press, 2014.

[5] Z. Tang, H. Meng, and D. Manocha, "Low-frequency compensated synthetic impulse responses for improved far-field speech recognition," in *ICASSP*. IEEE, 2020, pp. 6974–6978.

[6] C. Kim, A. Misra, K. 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 *INTERSPEECH*. ISCA, 2017, pp. 379–383.

[7] V. W. Neo, C. Evers, and P. A. Naylor, "Pevd-based speech enhancement in reverberant environments," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 186–190.

[8] T. Jenrungrot, V. Jayaram, S. M. Seitz, and I. Kemelmacher-Shlizerman, "The cone of silence: Speech separation by localization," in *NeurIPS*, 2020.

[9] C. Schissler, R. Mehra, and D. Manocha, "High-order diffraction and diffuse reflections for interactive sound propagation in large environments," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 1–12, 2014.

[10] A. H. Moore, M. Brookes, and P. A. Naylor, "Roomprints for forensic audio applications," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[11] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[12] Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha, "Improving reverberant speech training using diffuse acoustic simulation," in *ICASSP*. IEEE, 2020, pp. 6969–6973.

[13] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *ICLR*, 2019.

[14] S. Sakamoto, A. Ushiyama, and H. Nagatomo, "Numerical analysis of sound propagation in rooms using the finite difference time domain method," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3008–3008, 2006.

[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979. [Online]. Available: https://doi.org/10.1121/1.382599

[16] Z. Tang, J. D. Kanu, K. Hogan, and D. Manocha, "Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks," in *INTERSPEECH*. ISCA, 2019, pp. 654–658.

[17] C. Schissler and D. Manocha, "Interactive sound propagation and rendering for large multi-source scenes," *ACM Trans. Graph.*, vol. 36, no. 1, Sep. 2016. [Online]. Available: https://doi.org/10.1145/2943779

[18] A. Ratnarajah, Z. Tang, and D. Manocha, "Ts-rir: Translated synthetic room impulse responses for speech augmentation," *arXiv e-prints*, p. arXiv:2103.16804, Mar. 2021.

[19] M. R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *The Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 497–500, 1979.

[20] N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *The Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1484–1488, 1981.

[21] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio Engineering Society Convention 122*, May 2007. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14106

[22] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *ICLR*. OpenReview.net, 2018.

[23] Y. Li, S. Liu, J. Yang, and M. Yang, "Generative face completion," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5892–5900.

[24] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2242–2251. [Online]. Available: https://doi.org/10.1109/ICCV.2017.244

[25] H. Dong and Y. Yang, "Convolutional generative adversarial networks with binary neurons for polyphonic music generation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 190–196. [Online]. Available: http://ismir2018.ircam.fr/doc/pdfs/218_Paper.pdf

[26] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *NeurIPS*, 2019, pp. 14 881–14 892.

[27] N. J. Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," in *ICASSP*. IEEE, 2020, pp. 1–5.

[28] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[29] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *ICLR (Poster)*, 2016.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. X, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[31] 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

[32] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*. ISCA, 2015, pp. 3214–3218.

[33] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.