INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Blind Estimation of Room Impulse Response from Monaural Reverberant Speech with Segmental Generative Neural Network

Zhiheng Liao[1], *Feifei Xiong*[1], *Juan Luo*[1], *Minjie Cai*[2]
Eng Siong Chng[3], *Jinwei Feng*[1], *Xionghu Zhong*[3]

[1]Hummingbird Audio Lab, Alibaba Group, Hangzhou, China
[2]The University of Tokyo, Japan
[3]Nanyang Technological University, Singapore

Abstract

This paper presents a generative neural network to estimate room impulse response (RIR) directly from the received reverberant speech in single-channel scenario. Complex spectrogram of the reverberant speech is used as the input of an encoder to produce the compact acoustic embedding, which is then fed to a generator to construct the related time-domain acoustic response. To avoid a large model to generate the RIR with long taps, we propose SG-RIR, a novel segmental generative network that splits the RIR into segments and shares the network parameters across segments for blind RIR estimation. Experimental results show that the proposed model is capable of estimating the time-domain RIR with mean error of 0.008 in terms of both simulated and measured RIR test sets. The effectiveness is further verified by the achieved competitive estimation accuracy of two key room acoustic parameters (the reverberation time RT and the direct-to-reverberant ratio DRR) as compared to state-of-the-art approaches that are specific for RT and DRR estimation.

Index Terms: room impulse response, blind estimation, acoustic embedding, generative neural network

1. Introduction

The acoustic characteristics of a room have been shown to be important for many applications in audio analysis and speech processing, such as virtual sound in augmented reality audio [1], speech quality assessment [2], speech dereverberation [3] and distant speech recognition [4]. The room impulse response (RIR), representing the whole information of acoustic characteristics between two physical positions [5], can be measured manually, e.g., using an excitation signal such as a swept-sine signal [6]. However, RIR recordings require time and other resources, and are not always practical in real-world scenarios. Alternatively, acoustic simulators have been used for decades to generate synthetic/simulated RIRs using e.g., wave-based approach [7], or geometric-based approach with image source [8] or ray tracing [9]. On the other hand, their applications are limited due to the required geometric shape and material parameters of the target room environment, or too complex room geometry to simulate. Consequently, it is of great interest to blindly (or non-intrusively) estimate the real RIR. As direct estimate of the time-domain RIR from speech signals remains challenging due to its long taps (thousands of samples), research has been turned to estimate the key parameters of RIR, namely blind room acoustic parameter estimation. For instance, the reverberation time (RT), defined as the time interval for a 60 dB decay of the sound energy after the sound source is ceased, has drawn much attention for years [10, 11]. The direct-to-reverberant ratio (DRR) is another important parameter [12], referring to the energy ratio between the direct path and the reverberation caused by multi-path propagation from the sound source to the receiver. The summary of recent progress of blind room acoustic parameter estimation can be referred to [13]. While the estimated room parameters already provide useful knowledge of the acoustic characteristics, an auralization of the RIR is still more desirable: 1) though, given RT and DRR, RIRs can be modeled as white noise modulated by an exponentially decaying envelope [14, 10], such model can not capture subtleties such as early reflection patterns and coloration; 2) a complete time-domain RIR is generally required for augmented reality and reverb matching [15].

Recently, with the advances of deep learning, research on direct estimate of real time-domain RIR has been conducted. Improved generation of a realistic RIR has been proposed using neural networks, e.g., Ratnarajah et al. [16] proposed IR-GAN to generate more realistic synthetic RIRs with a generative adversarial network (GAN) fed by a latent vector drawn from a Gaussian distribution and be constrained by four key acoustic parameters. They further introduced TS-RIRGAN [17] to translate a simulated RIR to a more realistic RIR. FAST-RIR [18] focused on constructing specular and diffuse reflections in RIRs using conditional GAN with conventional environmental parameters. Improved RIR synthesis with multiple channels has been also studied in [19, 20]. In parallel, implicit RIR extraction in end-to-end neural framework has been introduced in acoustic matching [15] to transform speech recordings in a source environment to a target. The long taps of the real RIR still plays an obstructive role in explicit RIR estimation with neural networks. Alternatively, with the RIR modeling inspired from room acoustics [5] as a summation of decaying filtered noise signals along with the direct sound and early reflections, Steinmetz et al. [21] proposed FiNS that yields several orders of filtered noise signals to simulate late reverberation instead of estimating the common thousands of taps. Based on state-variable filter parameterization and frequency-sampling method, Lee at al. [22] introduced a differentiable artificial reverberation framework to estimate reverberation parameters for synthesizing RIR.

In an attempt to directly estimate the long but exact taps to explicitly construct the time-domain RIR and meanwhile, to avoid using a very large model, we propose a blind RIR estimation framework SG-RIR based on generative neural network with a novel segmental operation. More specifically, an encoder network is applied to extract the low-dimensional acoustic embedding from monaural reverberant speech, provided that such embedding could represent the acoustic characteristics only, i.e., invariant to speech content and speaker identity [15]. Then a generative network is adopted that is capable of constructing counterpart that is very similar to the real data, particularly incorporating a discriminator [23]. Inspired by subband net-

10.21437/Interspeech.2023-1344

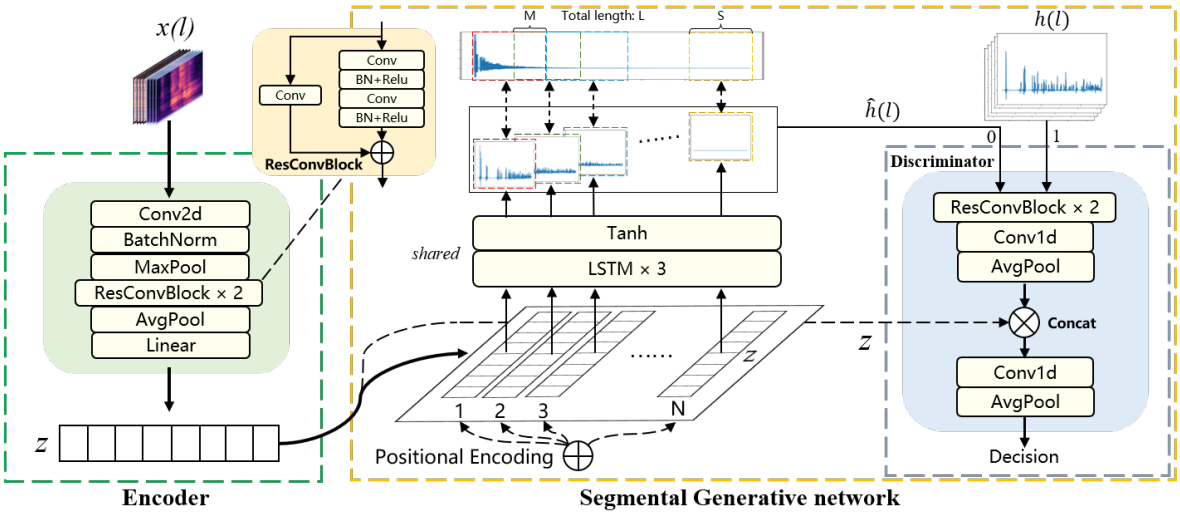

Figure 1: *The structure of SG-RIR consisting of an encoder and a segmental generative network, as well as a discriminator if applied.*

work [24, 25] in which network parameters can be shared across frequency bands, the proposed generative network is designed to yield a segmental component of the complete RIR each time with the repeated embedding input. Moreover, to label the sequence of the segments, positional encoding [26] is used and integrated into the embedding, and an overlap between the sequential RIR segments is introduced to eliminate the concatenation artifacts.

The remainder of this paper is organized as follows: We first introduce the novelty of the proposed method with the segmental network architecture and metrics-oriented loss function in Sec. 2, and the experimental setup for evaluation is described in Sec 3. Results and discussion are presented in Sec. 4 before Sec. 5 concludes the paper.

2. Proposed Method

The overall structure of the proposed SG-RIR is illustrated in Figure 1. First, the spectrograms of the reverberant speech signal $x(\ell)$, modeled by convolving anechoic speech to an RIR $h(\ell)$ with sample index ℓ, are fed to an encoder network to extract a latent acoustic embedding. The segmental generative network is then used to transform the acoustic embedding to the segmental RIR incorporating a discriminator if applied during training. The complete estimated RIR $\hat{h}(\ell)$ with length L is concatenated across the N sequential segments with overlap.

2.1. Encoder

The encoder follows the structure of ResNet introduced in [27], consisting of 2-dimensional convolution, batch normalization, max pooling layer, two ResConvBlocks, average pooling and linear layer. We use the spectrograms of the reverberant speech signal as input and evaluate the comparable performance between magnitude and complex (real and imaginary). A fixed Z-dimensional real latent embedding $z \in \mathbb{R}^Z$ is the output, which is expected to focus on capturing only the attributes of the RIR from the reverberant speech.

2.2. Segmental Generative Network

To avoid a very large output dimension of generative neural network, segmental generative network is designed to generate a segmental part of the RIR each time, and positional encoding [26] is adopted to label the sequence of the segments from 1 to N for the latter sequential concatenation. Inspired by the subband network [24] that is capable of sharing network parameters to yield independent output, the acoustic embedding is repeated along with the positional label to fed to a typical generative network, i.e., a stack of three long short-term memory (LSTM) layers with a Tanh activation to constrain the RIR range. Further, an overlap M between the sequential segments is introduced to avoid concatenate artifacts with segment length S. Considering the fixed-length L of the estimated RIR $\hat{h}(\ell)$, the amount of segments N is calculated as $\lfloor (L-M)/S \rfloor + 1$.

In addition, a discriminator can be employed to facilitate the training of the generative network, forming the GAN manner [28, 23, 29]. The estimated segmental time-domain RIR and the ground-truth counterpart are fed to the discriminator network, which is composed of two ResConvBlocks, 1-dimensional convolution and average pooling layer. In order to match the estimated segmental RIR with its ground-truth (to alleviate one-to-many issue), we further concatenate the embedding z as an additional label as illustrated in Figure 1, following another 1-dimensional convolution and average pooling layer before making the decision. As well, a discriminator can enable an unsupervised learning manner for the generative network, as mathematically expressed in the loss function as follows.

2.3. Loss Function

We use the signal-to-distortion ratio (SDR) loss as the minimization criterion to train the model with Adam optimizer, computed as

$$\mathcal{L}_{sdr} = \mathbb{E}\left\{ 10 \cdot \log_{10} \frac{||h(\ell) - \hat{h}(\ell)||^2}{||h(\ell)||^2} \right\}, \quad (1)$$

where $\mathbb{E}\{\cdot\}$ and $||\cdot||^2$ denote the expectation across all the segments and the L2 norm, respectively. Inspired by the use of evaluation metrics as loss function in [30], an auxiliary loss based on the derivable DRR computation [11] is introduced as

$$\mathcal{L}_{drr} = \mathbb{E}\left\{ \mathcal{F}_{drr}(h(\ell)) - \mathcal{F}_{drr}(\hat{h}(\ell)) \right\}, \quad (2)$$

$$\mathcal{F}_{drr}(Y(\ell)) = 10 \cdot \log_{10} \frac{\sum_{\ell=1}^{\ell=\ell_d} Y^2(\ell)}{\sum_{\ell=\ell_d+1}^{\ell=L} Y^2(\ell)}, \quad (3)$$

where $\mathcal{F}_{drr}(Y), Y = h \text{ or } \hat{h}$ denotes the DRR formula and ℓ_d represents the boundary sample separating the direct sound from RIR.

If the discriminator D is applied, the adversarial loss can be further exploited during training stage. Inspired by LSGAN [23] in which the loss function can stabilize the training, and CGAN [28] with conditional loss for one-to-one issue, we formulate the generator loss $\mathcal{L}_{gen}$ and the discriminator loss $\mathcal{L}_{dis}$ with the insertion of the embedding z as

$$\mathcal{L}_{gen} = \mathbb{E}\left\{(D(\hat{h}(\ell), z) - 1)^2\right\}, \quad (4)$$

$$\mathcal{L}_{dis} = \mathbb{E}\left\{(D(h(\ell), z) - 1)^2\right\} + \mathbb{E}\left\{(D(\hat{h}(\ell), z)^2\right\}. \quad (5)$$

As a result, the loss function for the proposed segmental generative network can be rewritten as

$$\mathcal{L}_{total} = \lambda_{sdr} \cdot \mathcal{L}_{sdr} + \lambda_{drr} \cdot \mathcal{L}_{drr} + \lambda_{gen} \cdot \mathcal{L}_{gen}, \quad (6)$$

where factor λ controls the contribution of respective loss function, e.g., $\lambda_{gen} = 0$ is the case without the discriminator, and $\lambda_{gen} \neq 0$ is the case with the discriminator and the loss $\mathcal{L}_{dis}$ in (5) is minimized simultaneously with minimizing $\mathcal{L}_{total}$, while $\lambda_{sdr}, \lambda_{drr} = 0$ means that the generative network turns to the unsupervised mode in terms of RIR generation.

3. Experimental Setup

The training data are synthesized using anechoic speech from TIMIT corpus [31] (training set with 6300 utterances) and a set of RIRs consisting of 1000 simulated and 940 real-measured RIRs. The simulated RIRs are generated using image method [8] with RT and DRR ranging from $0.3\,\mathrm{s}$ to $1.5\,\mathrm{s}$ and from $-15\,\mathrm{dB}$ to $10\,\mathrm{dB}$, respectively. The real-measured RIRs are collected from two open datasets including the Aachen Impulse Response datasets [32] and the OpenAir database [33], which cover 64 real scenarios with RT ranging from $0.1\,\mathrm{s}$ to $1.85\,\mathrm{s}$ and DRR from $-6\,\mathrm{dB}$ to $15\,\mathrm{dB}$. The training set is $47\,\mathrm{h}$ in total and all the utterances are sampled at $16\,\mathrm{kHz}$. To facilitate the network training, the input-target sequence pairs are set to a constant length, i.e., the reverberant speech signal is set to $2\,\mathrm{s}$ with the short-time Fourier transform (window length of 512 and shift length of 256 with Hann window) to obtain the spectrograms. The length of the estimated RIR is fixed to $L = 16384$ ($1.024\,\mathrm{s}$), and we set $S = 256, M = 128$ resulting in $N = 64$ (see Sec. 2.2). The embedding dimension Z is initially set to 128, and an ablation study will be carried out with different values (see Sec. 4.2). The boundary sample ℓ_d in (3) is chosen as the index of $2.5\,\mathrm{ms}$ according to [13]. The initialized learning rates for the encoder, the segmental generator and the discriminator are 2e-5, 4e-5, 1e-5, respectively, and we halve the rates when validation loss does not decrease until 100 epoch. The amount of the proposed segmental generative network parameters is 6.5 Million, and the real-time factor, calculated as the time of processing the complete RIR, is 0.09 on Intel Xeon CPU E5-2682 v4 (2.50 GHz) with Python implementation.

To guarantee non-overlap between training and test sets, the anechoic speech signals for test are taken from the TIMIT evaluation set, including 1960 utterances. In terms of simulated and real-measured RIRs, two test sets are created: 1) simulated test set with 14 simulated RIRs generated using image method but with different room parameters from the training set; 2) real test set with 14 measured RIRs from the ACE challenge test scenarios [13]. Each RIR is convolved with 200 (orderly selected) utterances from the TIMIT evaluation set, resulting in 2800 utterances ($3.2\,\mathrm{h}$) in each test set. Note that noises are not considered yet in this work which will be carried out in future work.

The estimation error $e_Y = \hat{Y} - Y$, i.e., the difference between the estimated value and the ground truth with Y denoting either the RIR, the RT or the DRR is used as evaluation metrics. The root mean squared error (RMSE) is reported for each measure ($\mathrm{RMSE}_{\mathrm{RIR}}$, $\mathrm{RMSE}_{\mathrm{RT}}$ and $\mathrm{RMSE}_{\mathrm{DRR}}$, respectively), as well as the Pearson correlation coefficients ρ_{RT}, ρ_{DRR} between estimated and true parameters (higher ρ towards 1 exhibiting more accurate estimates) as suggested in the ACE challenge [13].

4. Results

4.1. Overall Performance

As seen from the upper rows in Table 1 with simulated test set, when the auxiliary metrics-oriented loss function $\mathcal{L}_{drr}$ in (6) is applied ($\lambda_{drr} \neq 0$), performance of SG-RIR improves, particularly for the DRR estimation with nearly $1\,\mathrm{dB}$ decrease of $\mathrm{RMSE}_{\mathrm{DRR}}$. Generally speaking, the complex (real and imaginary) spectrogram as input provides better results compared to the magnitude counterpart, indicating the importance of the phase information required for the network to construct the time-domain RIR. Performance can be further improved with the attendance of the discriminator $\lambda_{gen} \neq 0$ to assist the training of the generator, particularly for the RT estimation. However, it is also clearly observed that SG-RIR performance degrades significantly when the generative network turns to the unsupervised mode (only with λ_{gen} in (6)). This indicates that solely depending on the GAN loss, it is challenging for SG-RIR to achieve an acceptable accurate RIR estimation, especially when the training data is not *big* enough (see Sec. 3).

Table 1: *SG-RIR Performance in terms of different spectrogram input and different λ in* (6) *w.r.t. the simulated (upper rows) and the real (lower rows) test set.*

Spectrogram	λ_{sdr}	λ_{drr}	λ_{gen}	RMSE ↓		
				RIR	RT(s)	DRR(dB)
magnitude	1.0	0.0	0.0	0.0082	0.226	6.874
magnitude	1.0	0.1	0.0	0.0080	0.225	5.910
complex	1.0	0.1	0.0	**0.0075**	0.248	3.985
complex	1.0	0.1	1.0	0.0076	**0.165**	**3.917**
complex	0.0	0.0	1.0	0.0780	0.654	11.215
complex	1.0	0.1	0.0	0.0083	0.255	4.747
complex	1.0	0.1	1.0	0.0087	0.176	3.985

Further, similar performance is achieved with the real test set (lower rows in Table 1), which hints at the generalization capabilities of the proposed SG-RIR for both simulated and real-measured RIRs. On average, it shows that SG-RIR can achieve the blind RIR estimation with time-domain amplitude error of 0.008, $\mathrm{RMSE}_{\mathrm{RT}}$ of $0.17\,\mathrm{s}$ and $\mathrm{RMSE}_{\mathrm{DRR}}$ of $3.95\,\mathrm{dB}$.

We also take two estimation examples to show the output of the proposed SG-RIR, as illustrated in Figure 2. It can be seen that SG-RIR generally produces similar RIR as the ground-truth in terms of the time-domain waveforms and the magnitude spectrograms. Compared to real-measured RIRs, it seems that SG-RIR provides more realistic results with simulated RIRs. This is more obvious with magnitude spectrograms where some irregular patterns in the real-measured RIR, especially at very low and very high frequency bands, can not be fully constructed. This

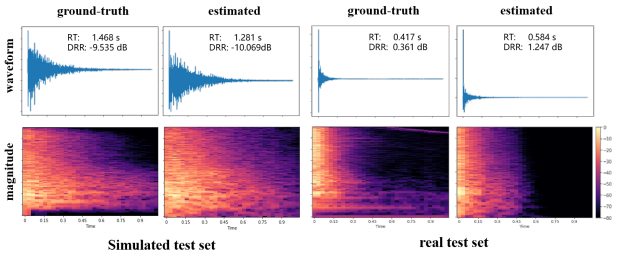

Figure 2: *Visualization of one simulated RIR and one measured RIR from the simulated and the real test set, respectively, as well as the estimated counterpart by SG-RIR.*

also partially explains the slightly superior performance with simulated test set in comparison to the real test set in Table 1. The interested readers are referred to a few more examples in https://github.com/ffxiong/sg-rir/.

4.2. Embedding Analysis

We further inspect the embedding z produced by the encoder (see Figure 1) with the simulated test set to analyze the underlying meaning of the learned vector. The UMAP tool [34] is used to project the high-dimensional embedding ($Z = 128$) into two-dimensional patterns for an easy visualization, as plotted in Figure 3. It demonstrates that clear cluster exists within the embedding space in terms of the ground-truth RT and DRR, indicating that the encoder has learned implicitly to capture the implicit knowledge about RIR based solely on the reverberant speech. This is also expected, as these cues are closely related to room acoustics for the RIR reconstruction and are invariant to the speech/phoneme content.

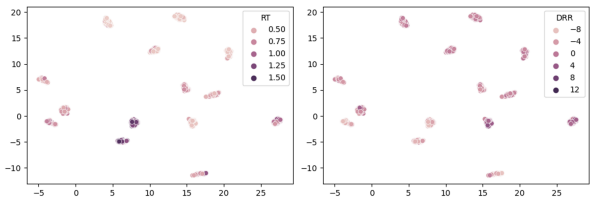

Figure 3: *Two-dimensional projections of the acoustic embedding with the simulated test set in terms of the ground-truth RT and DRR.*

Moreover, different dimensions Z of the acoustic embedding z are evaluated in terms of the averaged performance of SG-RIR w.r.t. both the simulated and the real test set, as shown in Figure 4. In general, performance improves as the dimension

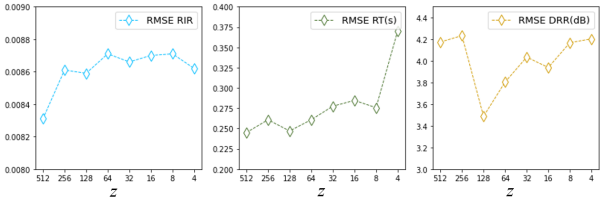

Figure 4: *Averaged performance of SG-RIR with both simulated and real test set in terms of different dimensions of the acoustic embedding.*

Z increases. On the other hand, performance tends to degrade for RT and DRR estimation when $Z > 128$, indicating that a non-compact embedding could hinder the generalization of the latter generator with the current network settings (see Sec. 3).

4.3. Performance Comparison with State-of-the-Art

To further verify the effectiveness of SG-RIR, two groups of state-of-the-art models are employed for performance comparison, one group includes the models specially designed for blind RT and DRR estimation, and the other group contains two models for blind time-domain RIR estimation. Note that all these models were implemented by their respective authors and tested on the same RIRs provided by the ACE challenge. Although the speech material for the test utterances is different and the noise scenarios are considered in the group of room parameter estimation, the performance could be still relatively comparable to some extent in terms of the RMSE and the correlation of RT and DRR metrics. As summarized in the upper rows in Table 2 with the first model group, results show that a fairly good RT estimator can achieve $\mathrm{RMSE}_{\mathrm{RT}}$ and ρ_{RT} of smaller than $0.25\,\mathrm{s}$ and larger than 0.70, respectively. Similarly, a fairly good DRR estimator can achieve $\mathrm{RMSE}_{\mathrm{DRR}}$ and ρ_{DRR} of smaller than $4.0\,\mathrm{dB}$ and larger than 0.55, respectively. With the values of RT and DRR directly derived from the estimated time-domain RIR, SG-RIR performance indicates that the proposed SG-RIR could be also considered as a good room parameter estimator. Moreover, in comparison to state-of-the-art RIR estimation models, better performance in terms of RMSE and comparable performance in terms of correlation can be achieved by SG-RIR for both RT and DRR metrics.

Table 2: *Performance comparison with other state-of-the-art models including specially designed room parameter estimators and RIR estimators w.r.t. the real test set consisting of the same real-measured RIRs.*

Model	$\mathrm{RMSE}_{\mathrm{RT}} \downarrow$	$\rho_{\mathrm{RT}} \uparrow$	$\mathrm{RMSE}_{\mathrm{DRR}} \downarrow$	$\rho_{\mathrm{DRR}} \uparrow$
QAReverb [35]	0.255	0.778	4.860	0.058
NIRA [12]	0.389	0.302	3.850	0.558
SRMR [36]	0.380	0.220	5.820	-0.084
ROPE [11]	0.285	0.716	4.810	0.556
jROPE [37]	0.288	0.758	4.090	0.621
Wave-U-Net [16]	0.367	0.324	7.019	0.681
FiNS [21]	0.237	0.837	6.605	0.640
SG-RIR	0.176	0.872	3.985	0.679

5. Conclusions

This paper proposes SG-RIR to accomplish blind estimation of the time-domain room impulse response given a monaural reverberant speech signal. By designing a novel segmental operation on the generative network and sharing the network parameters, our proposed model can directly produce the RIR with long taps. Experimental results show that the encoder is capable of extracting the compact acoustic embedding to represent the attributes of the RIR, and a lower estimation error is achieved by the segmental generator incorporating a metrics-inspired and an adversarial loss function. Comparable results with individual state-of-the-art approach of room acoustic parameter estimation and RIR estimation on a public test set further verify the effectiveness of the proposed framework. Future directions include the modeling of noisy scenarios, the integration of data augmentation, the subjective listening test, the lightweight network design and the online implementation for practical applications.

6. References

[1] J.-M. Jot, K. S. Lee, and E. Stein, "Augmented reality headphone environment rendering," in *AES Int. Conf. on Audio for Virtual and Augmented Reality*, Sep. 2016.

[2] P. C. Loizou, "Speech quality assessment," *Multimedia analysis, processing and communications*, pp. 623–654, 2011.

[3] K. Lebart, J.-M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[4] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, pp. 1–19, 2016.

[5] H. Kuttruff, *Room Acoustics*, 4th, Ed. London: Spon Press., 2000.

[6] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society 108th Convention*, Paris, France, Feb. 2000, pp. 1–23.

[7] 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, 2006.

[8] 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.

[9] C. Schissler and D. Manocha, "Interactive sound propagation and rendering for large multi-source scenes," *ACM Transactions on Graphics*, vol. 36, no. 1, 2016.

[10] R. Ratnam, D. L. Jones, B. C. Wheeler, J. W. D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[11] F. Xiong, S. Goetze, B. Kollmeier, and B. T. Meyer, "Exploring auditory-inspired acoustic features for room acoustic parameter estimation from monaural speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1809–1820, 2018.

[12] P. P. Parada, D. Sharma, T. van Waterschoot, and P. A. Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ACE challenge," in *IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, 2015.

[13] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[14] J. A. Moorer, "About this reverberation business," *Computer Music Journal*, pp. 13–28, 1979.

[15] J. Su, Z. Jin, and A. Finkelstein, "Acoustic matching by embedding impulse responses," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, Apr. 2020, pp. 426–430.

[16] A. Ratnarajah, Z. Tang, and D. Manocha, "IR-GAN: Room impulse response generator for far-field speech recognition," *Power*, vol. 140, pp. 286–290, 2021.

[17] ——, "TS-RIR: Translated synthetic room impulse responses for speech augmentation," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 259–266.

[18] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "FAST-RIR: Fast neural diffuse room impulse response generator," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 571–575.

[19] A. Richard, P. Dodds, and V. K. Ithapu, "Deep impulse responses: Estimating and parameterizing filters with deep networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3209–3213.

[20] M. Pezzoli, D. Perini, A. Bernardini, F. Borra, F. Antonacci, and A. Sarti, "Deep prior approach for room impulse response reconstruction," *Sensors*, vol. 22, no. 7, p. 2710, 2022.

[21] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 221–225.

[22] S. Lee, H.-S. Choi, and K. Lee, "Differentiable artificial reverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2541–2556, 2022.

[23] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[24] X. Li and R. Horaud, "Online monaural speech enhancement using delayed subband LSTM," in *Interspeech*, Sep. 2020, pp. 2462–2466.

[25] F. Xiong, W. Chen, P. Wang, X. Li, and J. Feng, "Spectro-temporal SubNet for real-time monaural speech denoising and dereverberation," in *Interspeech*, 2022, pp. 931–935.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communication ACM*, vol. 63, no. 11, pp. 139–144, oct 2020.

[30] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: an improved version of MetricGAN for speech enhancement," in *Interspeech*, Sep. 2021, pp. 201–205.

[31] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[32] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *16th International Conference on Digital Signal Processing*, 2009, pp. 1–5.

[33] D. T. Murphy and S. Shelley, "OpenAir: an interactive auralization web resource and database," in *Audio Engineering Society Convention 129*, 2010.

[34] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, 2018.

[35] T. d. M. Prego, A. A. de Lima, R. Zambrano-López, and S. L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," in *IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, 2015.

[36] M. Senoussaoui, J. F. Santos, and T. H. Falk, "SRMR variants for improved blind room acoustics characterization," in *IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, 2015.

[37] F. Xiong, S. Goetze, B. Kollmeier, and B. T. Meyer, "Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 255–267, 2019.