Generating Room Impulse Responses Using Neural Networks Trained with Weighted Combinations of Acoustic Parameter Loss Functions

Hualin Ren[1], Christian Ritz[1, *], Jiahong Zhao[2], Xiguang Zheng[1], Daeyoung Jang[3]
[1]School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, Australia
E-mail: hualin@uow.edu.au, E-mail: critz@uow.edu.au, E-mail: xiguang@uow.edu.au
[2]Institute of Sound and Vibration Research (ISVR), University of Southampton, Hampshire, UK
E-mail: Jiahong.Zhao@soton.ac.uk
[3]Media Coding Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea
E-mail: dyjang@etri.re.kr

***Abstract*— This paper investigates loss functions based on room acoustic parameters for training conditional generative adversarial networks (CGANs) used for generating room impulse responses (RIRs) at specific locations within a real room. The CGANs are trained on RIRs recorded at multiple positions in the room and then used to generate RIRs at new positions. The study evaluates the effectiveness of adaptive and fixed weightings, applied to various combinations of acoustic parameter-based loss functions, including those based on reverberation time (RT) and early decay time (EDT), along with the time-domain mean squared error (MSE) and frequency-domain multi-resolution short-time Fourier transform (MRSTFT). Results from reconstructing RIRs in two real rooms show that adaptive weightings are more effective than fixed weightings. MRSTFT surpasses MSE in accurately reconstructing RIRs across all frequency bands. Additionally, the loss functions using combinations of MRSTFT, RT30 and EDT with adaptive weightings significantly enhance the accuracy of RIR generation.**

I. INTRODUCTION

The room impulse response (RIR), which represents the acoustic characteristics of a room, is crucial in various audio signal processing tasks. For a realistic auditory experience in virtual reality (VR), the anechoic sound source must be continuously convolved with different RIRs as the listener's position changes within the virtual space. In streaming VR applications, providing an immersive experience in a real room with a complex shape by recording and transmitting RIRs for every possible listening position is impractical. Instead, RIRs can be modeled using the room's geometry, the reflection and absorption coefficients of its walls. However, this approach typically only produces satisfactory results for rooms with clear, simple geometries. Additionally, these methods often struggle to accurately model low frequencies and diffraction effects [1]. To address these limitations, artificial neural networks have been employed, to predict the desired data, offering a potential solution for generating RIRs in more complex environments.

Recently, generative adversarial networks (GAN) [2] based neural network approaches have been proposed for generating RIRs. GANs consist of two competing networks, a generator and a discriminator, which are trained in a back-and-forth manner to challenge each other. WaveGAN was the first attempt to generate raw-waveform audio using a GAN [3]. As an extension, a conditional generative adversarial net (CGAN) ensures that generated data match specific criteria [4]. Several CGAN-based models have been developed for RIR generation. Image2Reverb synthesizes RIRs using a CGAN with mean absolute error (MAE) loss and reverberant time 60 (RT60) losses [5]. FAST-RIR is a fast diffuse RIR generator employing mean squared error (MSE) and RT60 losses [6]. MESH2IR generates RIRs using MSE and energy decay relief losses [7].

However, the exploration of CGAN has only been partially explored in acoustics [8]. Existing RIR generators mainly focus on basic loss functions such as MAE and MSE [5], [6], with limited use of room acoustic parameters as loss functions. This study investigates the effects of some key room acoustic parameters, including RT and early decay time (EDT), on improving CGAN-based RIR generation, alongside the impact of fixed and adaptive weightings. These parameters are crucial for accurately reflecting acoustical properties of a room and enhancing the realism of the generated sound field [9], [10]. The hypothesis is that the acoustic characteristics of the RIR in a particular room can be learned, allowing for the generation of RIRs at arbitrary locations without needing detailed knowledge of the room geometry and surface reflection parameters.

Main Contributions: (1) Inclusion of acoustic parameters in loss functions. The study examines the use of key room acoustic parameters, such as RT and EDT, along with the time-domain MSE and frequency-domain multi-resolution short-time Fourier transform (MRSTFT) in loss functions for a CGAN-based network. (2) Investigation of the performance of

the fixed and adaptive weightings in the context of RIR generation. Fixed weightings remain constant throughout the training, but they might lead to imbalance weightings of loss functions. Adaptive weightings adjust with each iteration to balance the loss functions. (3) Uniquely integrating the MRSTFT with acoustic parameters in the loss functions, rather than using MRSTFT alone [11], [12], [13]. (4) Identification of the best-performing combination of all the mentioned acoustic characteristics in the loss functions for RIR generation. Evaluations compare recorded and generated RIRs using various metrics, including the MRSTFT, RT, EDT, direct-to-reverberant ratio (DRR), normalized root mean squared error (NRMSE), and DROQM for listening quality (LQ) [14].

II. CGAN-BASED RIR GENERATION

In this study, CGAN is trained to learn how to generate RIRs based on receiver and source positions in 3-dimensional Cartesian coordinates, as well as room dimensions, represented as a 9-dimensional embedding vector ε_p. The architecture of the CGAN model builds on the FAST-RIR model [6] and the Stage-I structure of StackGAN [15], with modifications including an additional upsample layer in the generator to accommodate the size of the input signal. The generator was organized by multiple transposed convolutional layers, while the discriminator receives a pair of real and generated RIRs as the input. The performance of the network is controlled by loss functions including MSE, MRSTFT, RT and EDT.

A. Generator Adversarial Loss

Conditioned on ε_p, the generator G is trained using a modified CGAN loss to minimize the overall error between the generated RIR $\tilde{\mathbf{p}}$ and the real-world recorded RIR $\mathbf{p}$. The discriminator, denoted by D, is used to distinguish between the generated RIRs and the real RIRs.

$$\mathcal{L}_{ADV}(G) = \mathbb{E}[(D(\tilde{\mathbf{p}}) - 1)^2] \tag{1}$$

B. MSE Loss

The MSE loss computes between each sample of the real-world recorded and generated RIRs. The $\|\cdot\|_{MSE}$ is the MSE loss between real and generated RIRs.

$$\mathcal{L}_{MSE}(\mathbf{p}, \tilde{\mathbf{p}}) = \| \mathbf{p} - \tilde{\mathbf{p}} \|_{MSE} \tag{2}$$

C. MRSTFT Loss

MRSTFT is extensively used in audio processing [11], [12], [13] as it effectively captures the time-frequency distribution of realistic waveforms [12]. It also prevents the generator from overfitting to a short-time Fourier transform (STFT) representation, which could lead to suboptimal performance in the waveform domain [11]. In this study, MRSTFT is calculated using four distinct STFT resolutions: 1024, 2048, 512, and 128. The hop size is set to half of the STFT length, to allow overlapping segments during transformation. A Hann window is applied in STFT, where $\|\cdot\|_F$ denotes the F-norm, and $|\mathfrak{F}\{\cdot\}|$ is the magnitude of a STFT operator that converts time-domain sound pressure vectors $\mathbf{p}(t)$ and $\tilde{\mathbf{p}}(t)$, into spectrograms of dimension $L \times W$, with L being the number of frequency bins and W the number of frames. The MRSTFT is computed as the sum of two terms, the spectral convergence $\mathcal{L}_{\mathrm{SC}}$ and the spectral log-magnitude $\mathcal{L}_{\mathrm{SM}}$ across these R different STFT resolutions.

$$\mathcal{L}_{SC}(\mathbf{p}, \tilde{\mathbf{p}}) = \frac{\| \, |\mathfrak{F}\{\mathbf{p}\}| - |\mathfrak{F}\{\tilde{\mathbf{p}}\}| \, \|_F}{\| \, |\mathfrak{F}\{\mathbf{p}\}| \, \|_F} \tag{3}$$

$$\mathcal{L}_{SM}(\mathbf{p}, \tilde{\mathbf{p}}) = \frac{1}{N} \| \, log|\mathfrak{F}\{\mathbf{p}\}| - log|\mathfrak{F}\{\tilde{\mathbf{p}}\}| \, \|_F \tag{4}$$

$$\mathcal{L}_{MRSTFT}(\mathbf{p}, \tilde{\mathbf{p}}) = \sum_{r=1}^{R} \mathcal{L}_{SC}(\mathbf{p}, \tilde{\mathbf{p}}) + \mathcal{L}_{SM}(\mathbf{p}, \tilde{\mathbf{p}}) \tag{5}$$

D. RT Loss

The RT loss is calculated in accordance with the standard ISO 3382-1:2009. In this paper, RT30 is chosen for the RT loss since the calculated RT60 sometimes exceeded the duration of the recorded RIRs in the selected databases.

$$\mathcal{L}_{RT}(\mathbf{p}, \tilde{\mathbf{p}}) = |\mathbf{p}_{\boldsymbol{RT30}} - \tilde{\mathbf{p}}_{RT30}| \tag{6}$$

E. EDT Loss

EDT is determined by fitting a straight line to the initial 10 decibels (dB) of sound decay, reflecting changes in perceived reverberance [9]. To predict reverberance ratings accurately, an average of EDT across frequency bands from 125 Hz to 2 kHz is calculated [16], following ISO 3382-1:2009 standard.

$$\mathcal{L}_{EDT}(\mathbf{p}, \tilde{\mathbf{p}}) = |\mathbf{p}_{\boldsymbol{EDT}} - \tilde{\mathbf{p}}_{EDT}| \tag{7}$$

F. Total Loss Function

An additional weighted loss term is used in the generator to ensure data consistency, represented by λ_{ADV}, λ_{MSE}, λ_{MRSTFT}, λ_{RT} and λ_{EDT}, respectively. A weight of 0 means this loss is not used to find the total loss while the values of non-zero weights determine the impact of this loss on the total loss.

The accuracy of RIR generation can be enhanced by incorporating additional auxiliary losses, such as acoustic parameters, alongside primary losses MRSTFT and MSE. This paper explores fixed and adaptive weightings. Fixed weightings are set as hyperparameters, remaining unchanged throughout training. However, they are often unsuitable for auxiliary tasks, as finding the appropriate weightings may require multiple training sessions. Without these adjustments, optimization imbalances can occur, where auxiliary losses either overshadow the main loss or are too weak to be effective.

To address this issue, adaptive weightings are applied based on the MetaBalance method [17]. This approach adjusts the gradient magnitude of auxiliary losses to better align with the target loss. By flexibly modifying the gradient of an auxiliary loss, it ensures a magnitude comparable to the target loss, promoting balanced optimization across multiple losses.

$$\mathcal{L}_{\mathrm{G}} = \lambda_{ADV}\mathcal{L}_{ADV} + \lambda_{MSE}\mathcal{L}_{MSE} + \lambda_{MRSTFT}\mathcal{L}_{MRSTFT} + \lambda_{RT}\mathcal{L}_{RT} + \lambda_{EDT}\mathcal{L}_{EDT} \tag{8}$$

$$\mathcal{L}_D = \mathbb{E}\left[(D(\mathbf{p}) - 1)^2 + \left(D(\tilde{\mathbf{p}})\right)^2\right] \tag{9}$$

Table 1. The evaluation of generated RIR by different loss functions in CUBE dataset.

Task	Loss function implementation	MRSTFT↓ (dB)	RT30↓ (s)	EDT↓ (s)	DRR↓ (dB)	NRMSE↓ (/)	DROQM↑ (LQ, DS, /)
1	MSE (main loss, fixed weightings)	4.38	0.38	0.56	3.21	2.74	0.45
2	MRSTFT (main loss, fixed weightings)	3.51	**0.01**	**0.04**	2.56	2.99	0.41
3	MSE (main loss, adaptive weightings)	4.38	0.24	0.48	3.60	2.88	0.49
4	MRSTFT (main loss, adaptive weightings)	**2.48**	0.29	0.20	**2.41**	**1.72**	**0.53**
5	Baseline (fixed weightings)	3.46	0.18	0.16	2.09	2.37	**0.47**
6	MRSTFT (main loss), RT30 (adaptive weightings)	2.78	**0.02**	0.07	2.19	2.41	0.44
7	MRSTFT (main loss), EDT (adaptive weightings)	2.95	0.04	0.09	2.28	2.58	0.43
8	MRSTFT (main loss), RT30, EDT (adaptive weightings)	**2.48**	**0.02**	**0.06**	**1.90**	**2.19**	0.45

Table 2. The evaluation of generated RIR by different loss functions in MeshRIR dataset.

Task	Loss function implementation	MRSTFT↓ (dB)	RT30↓ (s)	EDT↓ (s)	DRR↓ (dB)	NRMSE↓ (/)	DROQM↑ (LQ, DS, /)
1	MSE (main loss, fixed weightings)	2.71	0.03	0.08	0.86	1.68	0.51
2	MRSTFT (main loss, fixed weightings)	1.81	**0.01**	**0.02**	1.75	2.06	0.38
3	MSE (main loss, adaptive weightings)	5.62	0.25	0.30	3.32	2.45	0.45
4	MRSTFT (main loss, adaptive weightings)	**1.56**	0.06	0.05	**0.63**	**1.61**	**0.61**
5	Baseline (fixed weightings)	2.92	0.06	0.04	**0.68**	1.69	0.56
6	MRSTFT (main loss), RT30 (adaptive weightings)	1.47	0.07	0.12	2.00	1.58	**0.68**
7	MRSTFT (main loss), EDT (adaptive weightings)	1.47	**0.04**	0.04	0.75	1.71	0.55
8	MRSTFT (main loss), RT30, EDT (adaptive weightings)	**1.39**	**0.04**	**0.03**	**0.68**	**1.57**	0.63

* The bold and underlined text indicates the better performance for Table 1 and 2. The "/" means no unit.

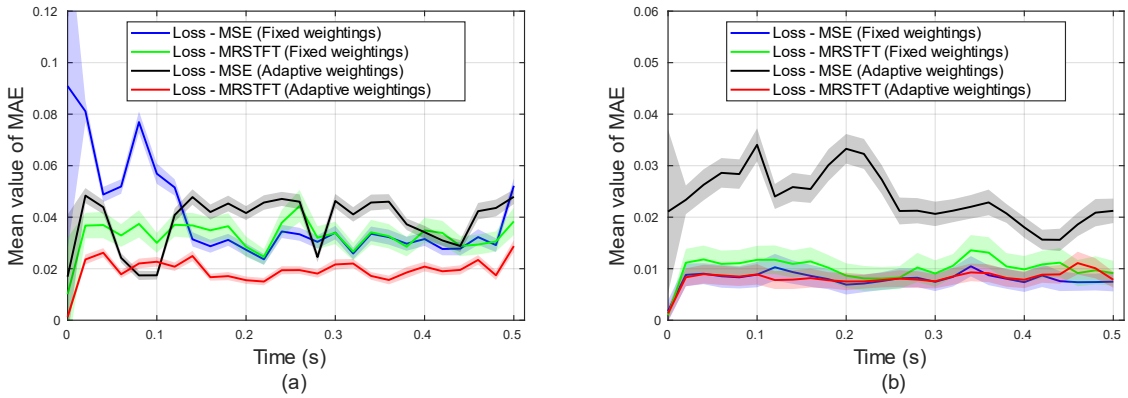

Fig. 1. Mean value of MAE with 95% confidence interval (shaded area) for fixed and adaptive weightings for MSE and MRSTFT as the main loss functions across time. (a) CUBE dataset. (b) MeshRIR dataset.

III. TRAINING

A. *Experimental Datasets*

The realistic RIR dataset CUBE was recorded in a studio measuring 10.5 m × 12 m × 5 m, with RT60 of 0.65 seconds (s) [18]. The dataset features 24 sources and 30 microphone positions, resulting in a total of 720 RIRs.

Another real recorded RIR dataset, MeshRIR, features a room with dimensions of 7 m × 6.4 m × 2.7 m and RT60 of 0.38 s [19]. For training and testing, the first, fifth, and last layers of the 3D cuboidal region (which has 9 layers) from the single source sub-dataset are selected, resulting in a total of 1323 RIRs.

Similar to [6], [13], all real-world recorded RIRs are resampled to a 16 kHz sampling rate and truncated to the first 8192 samples (0.512 s) to manage the resource limitations of the training server. Both datasets are divided into 80% for training and 20% for testing.

B. *Experimental Setup*

The CGAN model is trained on a GeForce RTX 3080Ti GPU for 800 epochs on both the MeshRIR and CUBE datasets. The optimization is performed with the RMSprop algorithm, a batch size of 128, and a learning rate of 8×10^{-5}. To minimize the impact of varying hyperparameters, all hyperparameters and network architecture are kept constant. The target output RIR

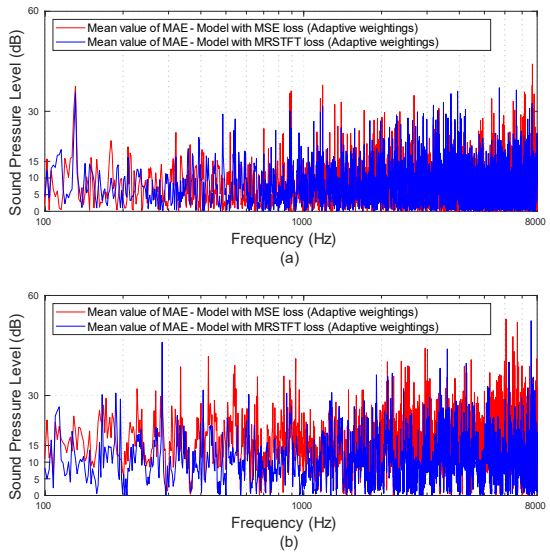

Fig. 2. Mean value of MAE of sound pressure level averaged over testing dataset for loss function MSE and MRSTFT. (a) CUBE dataset. (b) MeshRIR dataset.

is a 1D array of length 8192 samples at 16 kHz sampling rate, focusing on the direct sound (DS) and early reflection segments. The baseline model uses the loss functions from FAST-RIR [6], which include MSE and RT60 losses with fixed weightings. Eight experimental tasks are conducted to evaluate the performance of various combinations of loss functions. The tasks from 1 to 4 in both datasets (Table 1 and 2) investigate the weighting problems. The fixed weightings for the losses MSE, MRSTFT, RT30 and EDT are selected from [6] (λ_{ADV}= 1, λ_{MSE} and λ_{MRSTFT}= 0 or 20480, λ_{RT} and λ_{EDT}= 0 or 40), while adaptive weightings treat MSE and MRSTFT as main losses separately [17]. Tasks from 5 to 8 in Table 1 and 2 compare the performance of baseline, with models trained by the main loss MRSTFT and the acoustic parameters RT30 and EDT. Note that DRR was also considered as another potential loss function. However, during training, it was found that the convergence of this loss was unstable sometimes. Future work will investigate this further.

C. *Objective Evaluation*

The performance of the model is assessed using MRSTFT, RT30, EDT, DRR, NRMSE, and DROQM. DROQM is employed to evaluate the LQ (ranging from 0 to 1, where higher is better, with over 0.4 acceptable) of DS part (the most important part) of generated RIR with reference signal at the same position [14]. To analyze the effectiveness of MSE and MRSTFT in the frequency domain, the mean value of MAE of sound pressure level across the primary frequency range is calculated, as illustrated in Fig. 2, Table 3 and 4. In the time domain, the mean value of MAE across the entire signal is computed over short time windows of 20 milliseconds (ms), as

Table 3. Mean value of MAE of different frequency ranges averaged over testing datasets for loss function MSE and MRSTFT in CUBE dataset.

Frequency bands (Hz)	Loss function implementation	Mean value of MAE (dB)
100-500	MSE	**7.05**
	MRSTFT	7.16
501-1000	MSE	8.26
	MRSTFT	**7.57**
1001-4000	MSE	7.69
	MRSTFT	**7.00**
4000-8000	MSE	7.66
	MRSTFT	**7.06**

Table 4. Mean value of MAE of different frequency ranges averaged over testing datasets for loss function MSE and MRSTFT in MeshRIR dataset.

Frequency bands (Hz)	Loss function implementation	Mean value of MAE (dB)
100-500	MSE	15.84
	MRSTFT	**9.74**
501-1000	MSE	12.58
	MRSTFT	**10.16**
1001-4000	MSE	15.83
	MRSTFT	**10.63**
4000-8000	MSE	15.49
	MRSTFT	**10.59**

* The bold and underlined text indicates the better performance for Table 3 and 4.

shown in Fig. 1 and 3. Equation 10 details the MAE calculation, where $\mathbf{p}_m(k)$ denotes the measured sound pressure from M reference RIRs, and $\tilde{\mathbf{p}}_m(k)$ represents the estimated sound pressure of generated RIRs at various frequency points k.

$$Error_{MAE}(k) = \frac{1}{M}\sum_{m=1}^{M} |\mathbf{p}_m(k) - \tilde{\mathbf{p}}_m(k)| \quad (10)$$

IV. Results

A. *Fixed and Adaptive Weightings*

The findings demonstrate that adaptive weightings outperform fixed weightings. When using MRSTFT as the main loss function, adaptive weightings consistently achieve the lowest errors in MRSTFT, DRR, and NRMSE, as well as the highest DROQM scores across tasks from 1 to 4 in both datasets, as shown in Table 1 and 2. For the task focused on minimizing RT30 and EDT errors, the model with MRSTFT as the main loss function and fixed weightings proves more effective. These patterns are also evident in Fig. 1, where the model with MRSTFT loss and adaptive weightings (red line) generally outperforms others, achieving the lowest mean value of MAE. Although the MSE model with fixed weightings (blue line) shows the highest mean value of MAE from 0 to 0.1 s in Fig. 1(a), it performs comparably to the MRSTFT model with adaptive weightings in Fig. 1(b). This means that fixed weightings with MSE loss is not suitable for more reverberant room.

B. *MSE and MRSTFT Performance Comparison*

In tasks from 1 to 4, as presented in Table 1 and 2, models utilizing MRSTFT as the main loss function, whether with

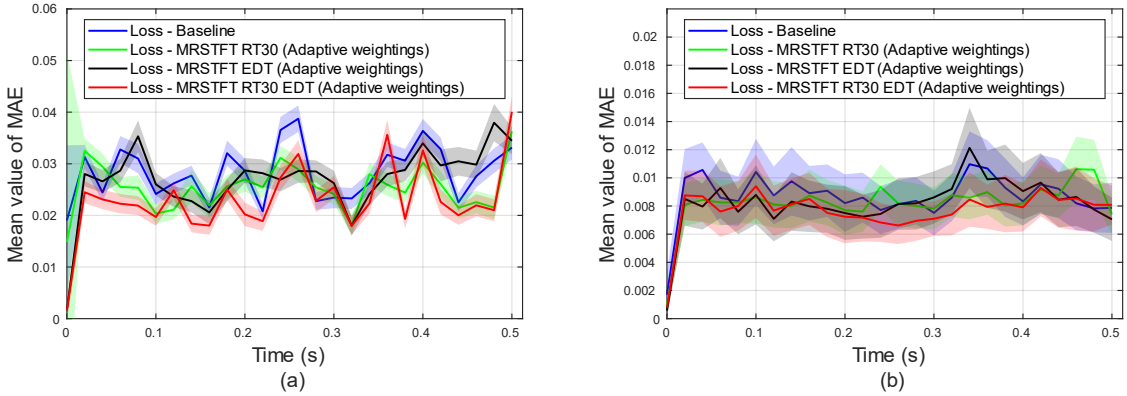

Fig. 3. Mean value of MAE with 95% confidence interval (shaded area) for different combinations of loss functions across time. (a) CUBE dataset. (b) MeshRIR dataset.

adaptive or fixed weightings, consistently achieve lower errors in most objective measurements compared to models using MSE as the main loss function. Additionally, the model using MSE as the main loss function with adaptive weightings (black line) performs the worst after 0.1 s in Fig. 1(a) and consistently underperforms in Fig. 1(b). As shown in Fig. 2(a) and 2(b), Table 3 and 4, the MRSTFT loss-based model demonstrates lower and more consistent mean value of MAE across all frequency bands. In contrast, the MSE loss-based model shows a higher and a wider range of mean value of MAE, as indicated in Table 3.

C. Baseline and Other Models Comparison

In Table 1 and 2, the model with MRSTFT, RT30, and EDT loss functions achieves the best performance across multiple objective measurements, including MRSTFT, RT30, EDT, DRR, and NRMSE. This combination outperforms both the baseline model and models using other loss functions. Specifically, the model with MRSTFT and RT30 loss functions results in the lowest RT30 error from Task 5 to 8 in Table 1, while the model with MRSTFT and EDT loss functions achieves the lowest RT30 error from the same tasks in Table 2. Considering the DS part simulation, the DROQM score is similar in more reverberant room in Table 1(CUBE dataset). However, in less reverberant room (Table 2, MeshRIR dataset), the models combining MRSTFT and RT30 losses, as well as those combining MRSTFT, RT30, and EDT losses, show better performance (top 2 scores in Table 1 and 2) in the DROQM metric. Fig. 3 supports these findings. The model with MRSTFT, RT30, and EDT losses (represented by the red line) consistently reaches the lowest error points, whereas the baseline model displays higher errors in most cases. Additionally, RIRs generated by models in the less reverberant room generally achieve lower MRSTFT, EDT, DRR, and NRMSE errors, along with higher DROQM scores, as shown from Task 5 to 8 in Table 1 and 2, and display more stability in Fig. 3(a) and 3(b), indicating more accurate generation in the less reverberant environment.

V. Discussion

The MRSTFT loss function consistently outperforms the MSE loss function, demonstrating superior performance in both frequency domain (as indicated by MRSTFT metric) and time domain (as shown by NRMSE metric). Specifically, the MRSTFT enhances the recovery of RIR energy across all frequency bands, particularly at higher frequencies. Additionally, incorporating various combinations of acoustic parameters in the loss functions significantly improves the performance compared to that of the baseline model. In experiments using both datasets, the loss function combining MRSTFT, RT30, and EDT losses with adaptive weightings demonstrates better performance, achieving the lowest errors in most cases. The effectiveness of the room acoustics parameters varies with different room conditions. In the more reverberant room, the differences in errors are more observable, with variations sometimes reaching a factor of ten, making the impact of different combinations of the loss functions more detectable. Conversely, in the less reverberant room, the errors are closer to each other. Overall, RIR generation performs more accurately in the less reverberant room.

VI. Conclusions

This study evaluates the effectiveness of fixed and adaptive weightings, compares the performance of MSE and MRSTFT, and explores various combinations of loss functions with key acoustic parameters in CGAN-based RIR generators, focusing on their ability to accurately synthesize RIRs. The findings demonstrate that adaptive weightings outperform fixed weightings. The MRSTFT shows significant advantages over the MSE, enhancing the performance of RIR generation in both the frequency and time domains. Notably, highly accurate results are obtained by incorporating adaptive weightings with the MRSTFT, acoustic parameters RT30 and EDT as loss functions. This innovative integration offers a new perspective on improving RIR generation. Further explorations of loss functions could lead to even more precise synthesized RIRs and then their applications such as higher-order Ambisonics.

VII. ACKNOWLEDGMENT

This research was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government, grant number 23ZH1200 (The research of the basic media contents technologies).

REFERENCES

[1] M. Tamulionis, T. Sledevič, and A. Serackis, "Investigation of Machine Learning Model Flexibility for Automatic Application of Reverberation Effect on Audio Signal," *Applied Sciences*, vol. 13, no. 9, Art. no. 9, Jan. 2023, doi: 10.3390/app13095604.

[2] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, in NIPS'14. Cambridge, MA, USA: MIT Press, Dec. 2014, pp. 2672–2680.

[3] C. Donahue, J. McAuley, and M. Puckette, "Adversarial Audio Synthesis," presented at the International Conference on Learning Representations, Feb. 2018.

[4] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv.org. Accessed: Sep. 02, 2023. [Online]. Available: https://arxiv.org/abs/1411.1784v1

[5] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori, "Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 286–295. doi: 10.1109/ICCV48922.2021.00035.

[6] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "Fast-Rir: Fast Neural Diffuse Room Impulse Response Generator," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 571–575. doi: 10.1109/ICASSP43922.2022.9747846.

[7] A. Ratnarajah, Z. Tang, R. Aralikatti, and D. Manocha, "MESH2IR: Neural Acoustic Impulse Response Generator for Complex 3D Scenes," in *Proceedings of the 30th ACM International Conference on Multimedia*, Lisboa Portugal: ACM, Oct. 2022, pp. 924–933. doi: 10.1145/3503161.3548253.

[8] P. Gerstoft, H. Groll, and C. F. Mecklenbräuker, "Parametric Bootstrapping of Array Data with A Generative Adversarial Network," in *2020 IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Jun. 2020, pp. 1–5. doi: 10.1109/SAM48682.2020.9104371.

[9] J. S. Bradley, "Review of objective room acoustics measures and future needs," *Applied Acoustics*, vol. 72, no. 10, pp. 713–720, Oct. 2011, doi: 10.1016/j.apacoust.2011.04.004.

[10] S. Saini, I. Engel, and J. Peissig, "An end-to-end approach for blindly rendering a virtual sound source in an audio augmented reality environment," *J AUDIO SPEECH MUSIC PROC.*, vol.2024, no. 1, p. 16, Mar. 2024, doi: 10.1186/s13636-024-00338-6

[11] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6199–6203. doi: 10.1109/ICASSP40776.2020.9053795.

[12] S. Joshi *et al.*, "Defense against Adversarial Attacks on Hybrid Speech Recognition System using Adversarial Fine-tuning with Denoiser," Sep. 2022, pp. 5035–5039. doi: 10.21437/Interspeech.2022-10977.

[13] E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstoft, "Generative models for sound field reconstruction," *The Journal of the Acoustical Society of America*, vol. 153, no. 2, pp. 1179–1190, Feb. 2023, doi: 10.1121/10.0016896.

[14] H. Ren, C. Ritz, J. Zhao, and D. Jang, "Towards an Objective Quality Metric for Interpolated Directional Room Impulse Responses," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 8205–8209. doi: 10.1109/ICASSP48485.2024.10446507.

[15] H. Zhang *et al.*, "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5908–5916. doi: 10.1109/ICCV.2017.629.

[16] M. Barron, "Subjective Study of British Symphony Concert Halls," *Acta Acustica united with Acustica*, vol. 66, no. 1, pp. 1–14, Jun. 1988.

[17] Y. He, X. Feng, C. Cheng, G. Ji, Y. Guo, and J. Caverlee, "MetaBalance: Improving Multi-Task Recommendations via Adapting Gradient Magnitudes of Auxiliary Tasks," in *Proceedings of the ACM Web Conference 2022*, in WWW '22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 2205–2215. doi: 10.1145/3485447.3512093.

[18] K. Müller, "CUBE B-format RIR dataset (Soundfield ST450 MKII)". Available: https://phaidra.kug.ac.at/o:104435

[19] S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, and J. Brunnström, "MESHRIR: A Dataset of Room Impulse Responses on Meshed Grid Points for Evaluating Sound Field Analysis and Synthesis Methods," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2021, pp. 1–5. doi: 10.1109/WASPAA52581.2021.9632672.