2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Towards a B-format Ambisonic Room Impulse Response Generator Using Conditional Generative Adversarial Network

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 proposes a B-format Ambisonic room impulse response (RIR) generator based on a conditional generative adversarial network (CGAN). The B-format RIR is a first-order Ambisonics (FOA) representation of the RIR used for creating spatial audio for virtual reality (VR) applications. The neural network produces FOA RIRs for specific VR rooms based on given receiver and source positions, and room dimension. The CGAN is trained using real-world FOA RIR recordings, with generative adversarial loss and other loss functions. These loss functions include the frequency-domain multi-resolution short-time Fourier transform (MRSTFT) loss function and various acoustic parameters, such as reverberation time (RT), early decay time (EDT), and direct-to-reverberant ratio (DRR). Adaptive weightings are applied to balance the importance of different loss functions. Objective evaluation metrics include mean squared error, MRSTFT, RT, EDT, DRR, DROQM for listening quality and direction of arrival accuracy. The results show that utilizing acoustic parameters as loss functions effectively controls these metrics, resulting in high-quality FOA RIRs.**

I. Introduction

In the application of virtual reality (VR), spatial audio is critical to ensure an immersive listening experience. Simulating spatial audio within rooms requires convolution of the sound source with the room impulse response (RIR) between the receiver and source positions. A first-order Ambisonic (FOA) RIR is a B-format array recording. The FOA RIR is beneficial for rendering spatial audio in VR environments corresponding to a listener's changing position and look direction relative to the source, known as six degrees of freedom spatial audio.

For streaming VR applications, it can be impractical to record and transmit the RIR for every possible listening position in a virtual room. Alternatively, RIRs for the chosen source and receiver positions can be synthesized based on acoustic signal processing models that utilize knowledge of the room geometry and wall reflection and absorption coefficients. The well-known image method [1] works well for shoebox room models while more complex methods are required for rooms with irregular geometries and containing reflective surfaces other than walls (e.g. furniture). However, these simulators can be too computationally complex to quickly update the RIR to match a listener's moving position in a virtual room. Alternatively, interpolation approaches applied to the FOA RIR can be used to produce an interpolated FOA RIR at a new position [2], [3], [4], [5] using the sound source direction information derived from the FOA RIR channel relationships. While promising, these approaches still require regular transmission of two or more FOA RIRs to facilitate the interpolation of new FOA RIRs matching the listener's position. In addition, accurately simulating the specular and diffuse reflections for a given room is difficult for conventional FOA RIR synthesis.

Recently, neural network approaches have been proposed for generating RIRs corresponding to arbitrary receiver and source positions in a room. Generative adversarial networks (GAN) [6] are widely applied in image generators, composed of two models, generator G and discriminator D, trained in a back-and-forth manner to challenge each other. WaveGAN is the first attempt to generate raw-waveform audio based on GAN [7]. As an extension, IR-GAN is proposed as a WaveGAN based RIR generator [8]. However, [8] does not control the acoustic characteristics of RIR (e.g. reverberation time 60 (RT60)) during the training session. Conditional generative adversarial net (CGAN) is introduced as an extension of GAN, to ensure that generated data (e.g. images) is constrained according to match certain desirable criteria [9]. For example, Image2Reverb synthesizes RIRs using a CGAN using images of acoustic environments as input [10]; FAST-RIR is a fast diffuse RIR generator with the inputs including receiver positions, source positions, room dimensions, and room acoustics parameters [11]; and MESH2IR generates RIRs using a three-dimensional (3D) scene mesh of the room, based on the FAST-RIR architecture [12]. These methods focus on improving the augmentation of far-field automatic speech recognition. Considering RIR interpolation on a spatial grid application, SoundNeRirF adopts multi-layer perceptrons (MLPs) to predict a forward RIR for the target position based on the 3D position of reference and target receivers [13]. Existing approaches aim to generate a mono RIR, while the proposed approach fills the gap to generate FOA RIRs.

Main Contributions: (1) The CGAN-based neural FOA RIR generator is introduced to generate plausible FOA RIRs at the chosen receiver position, source position and room dimension in the given acoustic environment using various combinations of loss functions. Without transmitting the known FOA RIRs, this approach improves the transmission efficiency by simply inputting the receiver position and source position to generate a new FOA RIR. (2) Frequency-domain multi-resolution short-time Fourier transform (MRSTFT) and multiple key acoustic parameters including RT30, early decay time (EDT), and direct-to-reverberant ratio (DRR) are investigated to find the best performance of losses combination and acoustic parameter losses. (3) Adaptive weightings are implemented to balance the main loss and auxiliary losses. (4) The model is trained using recordings of FOA RIRs from a real-world room, while existing literatures in this field mostly work on synthetic mono RIR datasets only or their synthetic mixture with real-world recorded datasets [8], [11], [12], [13]. Evaluations compare the difference between the recorded FOA RIRs and the generated FOA RIRs at the same pair of receiver positions, source positions and room dimensions using measures of mean squared error (MSE), MRSTFT, RT30, EDT and DRR, along with DROQM (for listening quality (LQ)) [14] and direction of arrival (DOA) accuracy of direct sound (DS) [15].

II. Proposed System

The GAN has been shown to generate realistic audio waveforms [7]. Here, a CGAN is used as the basis for the proposed generator to learn the specular and diffuse reflections. The proposed network takes the receiver position, source position and room dimension, represented in 3D Cartesian coordinates (x_r, y_r, z_r), (x_s, y_s, z_s) and (l_r, w_r, h_r), as the vector embedding $\varepsilon_p = (x_r, y_r, z_r, x_s, y_s, z_s, l_r, w_r, h_r)$. Each vector embedding is obtained directly from the real-world recorded dataset. The output is one-dimensional (1D) raw-waveform FOA RIR with four channels in Ambisonic channel number (ACN) order *W*, *Y*, *Z* and *X*. The sampling frequency of the generated FOA RIR is 16 kHz, with the length being 8192. The architecture of the proposed model is modified from FAST-RIR [11] that is based on the Stage-I structure of StackGAN [16]. Input of the proposed network is the vector embedding ε_p. The target real-world recorded FOA RIR is the 1D signal with four channels, similar to image inputs with RGB channels. The size of an input signal is set to (32, 4, 8192), corresponding to batch size, channel and height, so that the Dataloader is capable of inputting and outputting 1D FOA RIR (4, 8192), 4 arrays with each length being 8192. Another upsample layer of generator is also added, compared to the FAST-RIR architecture, to adapt the size of the input signal. To generate FOA RIRs in any specific position, the proposed network controls loss functions including the MRSTFT, RT, DRR and EDT.

A. Generator Modified CGAN Loss

Conditioned on ε_p, the generator G is trained through modified CGAN loss to minimise the overall error between the generated FOA RIR and real-world recorded FOA RIR. The D represents the discriminator, while the generated FOA RIR is $\tilde{\mathbf{p}}$ and the real-world recorded FOA RIR is $\mathbf{p}$.

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

B. MRSTFT Loss

The multi-resolution short-time Fourier transform (MRSTFT) is widely utilized in audio processing [17], [18], [19] due to its effectiveness in capturing the time-frequency distribution of realistic waveforms [18]. This approach helps avoid the generator's overfitting to a single short-time Fourier transform (STFT) representation, which can otherwise degrade performance in the waveform domain [17]. In this study, MRSTFT is computed using four different STFT resolutions: 1024, 2048, 512, and 128. The hop size is set to half the length of the STFT to enable overlapping segments during the transformation process. A Hann window is applied in STFT, where $|\mathfrak{F}\{\cdot\}|$ 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. The MRSTFT loss is calculated as the sum of two terms: spectral convergence $\mathcal{L}_{\mathrm{SC}}$ and spectral log-magnitude $\mathcal{L}_{\mathrm{SM}}$ across the various STFT resolutions.

$$\mathcal{L}_{SC}(\mathbf{p}, \tilde{\mathbf{p}}) = \frac{\| \, |\mathfrak{F}\{\mathbf{p}\}| - |\mathfrak{F}\{\tilde{\mathbf{p}}\}| \, \|_F}{\| \, |\mathfrak{F}\{\mathbf{p}\}| \, \|_F} \qquad (2)$$

$$\mathcal{L}_{SM}(\mathbf{p}, \tilde{\mathbf{p}}) = \frac{1}{N} \| \, log|\mathfrak{F}\{\mathbf{p}\}| - log|\mathfrak{F}\{\tilde{\mathbf{p}}\}| \, \|_F \qquad (3)$$

$$\mathcal{L}_{MRSTFT}(\mathbf{p}, \tilde{\mathbf{p}}) = \sum_{c=1}^{4} \mathcal{L}_{SC}(\mathbf{p}, \tilde{\mathbf{p}}) + \mathcal{L}_{SM}(\mathbf{p}, \tilde{\mathbf{p}}) \qquad (4)$$

C. RT Loss

The RT loss is based on the standard ISO 3382-1:2009. The absolute loss is computed on each channel, and the total loss is added together. RT30 is calculated as the time it takes for the sound level to decay by 30 decibels (dB) after the sound source has stopped emitting.

$$\mathcal{L}_{RT30}(\mathbf{p}, \tilde{\mathbf{p}}) = \sum_{c=1}^{4} |\mathbf{p}_{\boldsymbol{RT30}} - \tilde{\mathbf{p}}_{RT30}| \qquad (5)$$

D. DRR Loss

The DRR measures the energy ratio of sound arriving directly from a source to that arriving after one or more reflections from surfaces [20]. To compute the DRR, the direct impulse is first identified, and then the direct sound (DS) component is calculated by measuring the sound energy within 5 milliseconds (ms) around the direct impulse [21].

$$\mathcal{L}_{DRR}(\mathbf{p}, \widetilde{\mathbf{p}}) = \sum_{c=1}^{4} |\mathbf{p}_{\boldsymbol{DRR}} - \widetilde{\mathbf{p}}_{\boldsymbol{DRR}}| \tag{6}$$

E. EDT Loss

EDT is determined by fitting a straight line to the initial 10 dB of sound decay, indicating changes in perceived reverberance [22]. To predict reverberance ratings accurately, the average EDT is calculated across frequency bands from 125 Hz to 2 kHz [23], in accordance with ISO 3382-1:2009 standard.

$$\mathcal{L}_{EDT}(\mathbf{p}, \widetilde{\mathbf{p}}) = \sum_{c=1}^{4} |\mathbf{p}_{\boldsymbol{EDT}} - \widetilde{\mathbf{p}}_{\boldsymbol{EDT}}| \tag{7}$$

F. Total Loss Function

The total loss function for the FOA RIR generation integrates several distinct loss components to achieve the most accurate results. The weights assigned to the CGAN loss, MRSTFT, RT30, EDT, DRR losses are controlled by parameters λ_{ADV}, λ_{MRSTFT}, λ_{RT}, λ_{DRR} and λ_{EDT}, respectively. A weight of 0 excludes a loss function from the total loss, while non-zero weights determine the influence of each loss function.

To enhance the accuracy of RIR generation, additional auxiliary losses, such as acoustic parameters, are incorporated alongside primary loss MRSTFT. This paper investigates the use of adaptive weightings, as opposed to fixed weightings commonly used in prior work [11], which are set as hyperparameters and remain unchanged throughout training. Fixed weightings can be inadequate for managing multiple loss functions, often requiring numerous training processes to find the optimal values. In the absence of adaptive adjustments, there can be imbalances where auxiliary losses either dominate or fail to contribute effectively. To address this, adaptive weightings are employed based on the MetaBalance method [24]. This approach adjusts the gradient magnitudes of auxiliary losses (RT30, EDT and DRR losses) to better align with the target loss (MRSTFT loss), ensuring a balanced optimization process by maintaining comparable magnitudes for all losses involved.

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

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

III. Experiments

The proposed model is trained via a GeForce RTX 3080Ti GPU over 600 epochs on the dataset demonstrated below. The optimizer is RMSprop, and the learning rate is 2×10^{-4}.

A. Datasets

The realistic FOA RIR dataset is recorded by using B-format microphones in a room of the Institute of Electronic Music and Acoustics in Graz, Austria [25]. The FOA RIR is carried out in a 10.5 m × 12 m × 5 m studio in ACN order before being normalized. The number of FOA RIRs is 720, while the whole dataset is split into training and testing datasets, with their percentages being 90% and 10%, respectively. Similar to [11], the real-world recorded FOA RIRs are resampled to 16 kHz, and 9600 samples is the length, i.e. 0.6 s. In the experiments, the recorded FOA RIR length is truncated to 8192 samples, corresponding to 0.512 s. The FOA RIRs are normalized by the maximum absolute value among the four channels to maintain the relationships within the channels.

B. Implementations Evaluated

Eight experiments are conducted to assess the performance of the proposed network with various loss combinations. Each experiment builds upon the base CGAN loss described in equation (1). The additional combinations tested include: CGAN loss combined with MRSTFT loss in (4); CGAN loss combined with MRSTFT and RT30 losses in (5), DRR losses in (6), EDT losses in (7); and CGAN loss combined with MRSTFT loss and various combinations of RT30, DRR and EDT losses in (8).

C. Objective Evaluation

The objective metrics are used to evaluate performance by several key measures: MSE over different lengths, MRSTFT, RT30, DRR, EDT, DOA, and DROQM. These metrics assess the impact of various loss functions on the generated FOA RIRs. Errors for MRSTFT, RT30, EDT, and DRR are calculated consistently with their corresponding loss functions, ensuring uniformity in evaluation. The MSE metric is computed over various lengths, including the entire signal, 5 ms and 50 ms for DS length. DROQM predicts the impact on LQ (0 to 1, where higher is better, with over 0.4 acceptable) of the DS part of FOA RIRs [14]. The errors presented in Table 1 are the average absolute errors between the real-world recorded FOA RIRs and the generated FOA RIRs averaged across all four channels for each pair of receiver and source positions in the testing dataset.

The DOA information is extracted from FOA RIR. The four channels W, X, Y and Z are derived from a B-format microphone positioned at a specific azimuth angle and elevation angle. The omnidirectional channel W represents the impulse response of the audio, while the three orthogonal directional components X, Y, and Z channels capture directional cues. After converting the signal to the time frequency domain using the STFT, the azimuthal DOA is estimated for each time-frequency, where n is the frame number and k is the frequency index, as [15],

$$p_w(n,k) = \frac{W(n,k) \times conj(W(n,k))}{\sum_{n,k}^{N,K} W(n,k) \times conj\left(W(n,k)\right)} \tag{10}$$

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

Loss function implementation	**MSE↓ (DS, 5 ms, /)**	**MSE↓ (DS, 50 ms, /)**	**MRSTFT↓ (dB)**	**RT↓ (s)**	**EDT↓ (s)**	**DRR↓ (dB)**	**DOA↓ Azimuth (DS, radian)**	**DOA↓ Elevation (DS, radian)**	**DROQM↑ (DS, LQ, /)**
MRSTFT	0.04	**0.01**	3.14	0.20	0.19	2.64	0.82	0.42	0.44
MRSTFT, RT30	0.03	0.02	3.44	0.16	0.12	2.75	0.80	0.48	0.45
MRSTFT, EDT	**0.02**	**0.01**	3.30	0.21	0.14	2.28	**0.69**	**0.36**	0.43
MRSTFT, DRR	0.04	**0.01**	3.32	0.23	0.19	2.20	0.84	0.45	0.42
MRSTFT, RT30, EDT	0.04	0.02	3.67	0.24	0.25	2.29	0.81	0.53	**0.46**
MRSTFT, RT30, DRR	0.03	**0.01**	**3.10**	0.16	0.16	2.61	0.78	0.47	**0.46**
MRSTFT, EDT, DRR	0.04	**0.01**	3.21	0.18	0.11	2.43	0.89	0.41	0.41
MRSTFT, RT30, EDT, DRR	0.05	0.02	3.42	**0.15**	**0.10**	**1.96**	0.78	0.43	0.41

* The bold and underlined text indicates the lowest error in Table 1. The "/" means no unit.

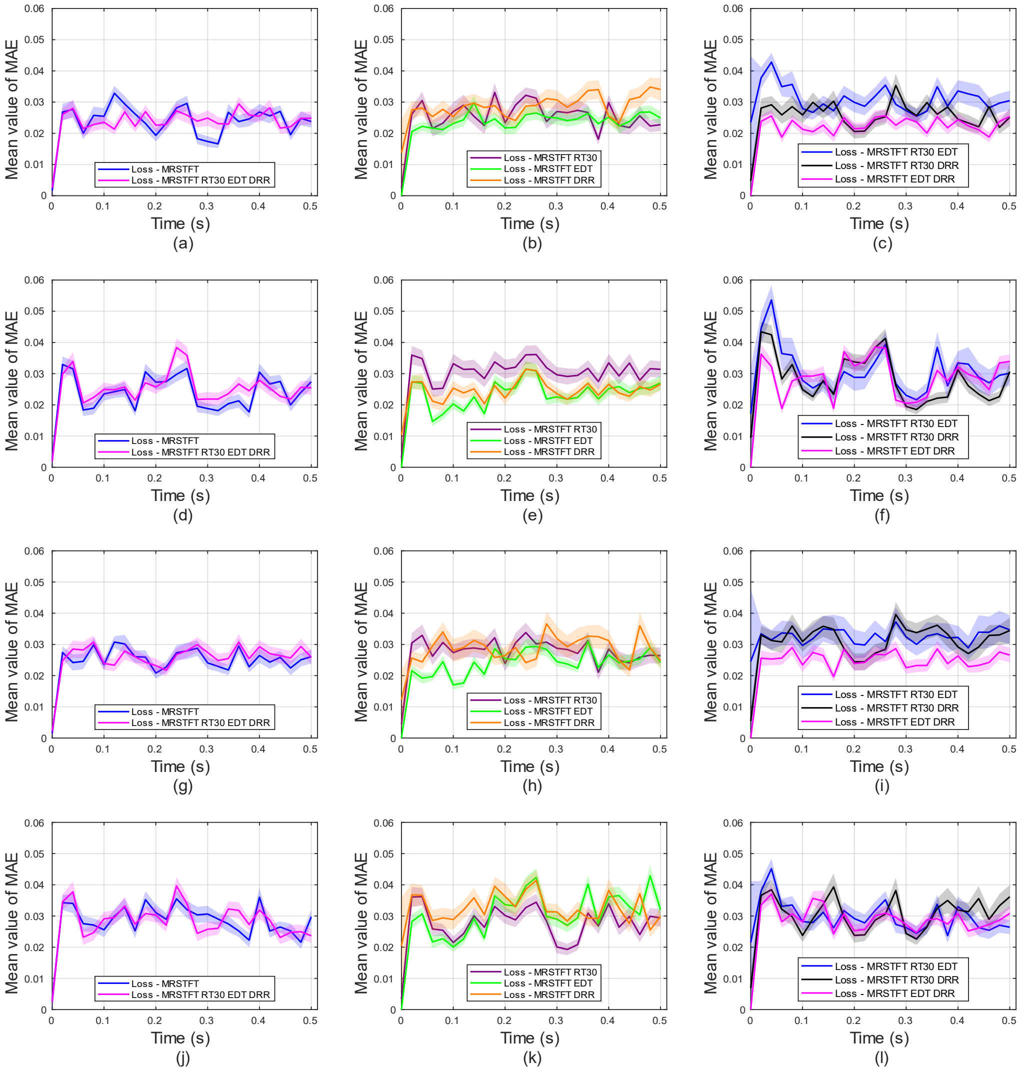

Fig. 1. Mean value of MAE with 95% confidence interval (shaded area) for the model using MRSTFT as the main loss function, combined with various acoustic parameters as auxiliary loss functions across time. (a) - (c) for W channel, (d) - (f) for Y channel, (g) - (i) for Z channel, (j) - (l) for X channel

$$I_{C=X,Y,Z}(n,k) = Re(conj(W(n,k)) \times C(n,k)) \quad (11)$$

$$I_{azi}(n,k) = \frac{I_{C=Y}(n,k)}{I_{C=X}(n,k)} \quad (12)$$

$$I_{ele}(n,k) = \frac{-I_{C=Z}(n,k)}{\sqrt{\left(I_{C=X}(n,k)\right)^2 + \left(I_{C=Y}(n,k)\right)^2}} \quad (13)$$

$$DOA_{azi} = \sum_{n,k}^{N,K} p_w(n,k) \times Re\left(\tan^{-1}\left(I_{azi}(n,k)\right)\right) \quad (14)$$

$$DOA_{ele} = \sum_{n,k}^{N,K} p_w(n,k) \times Re\left(\tan^{-1}\left(I_{ele}(n,k)\right)\right) \quad (15)$$

where, $p_w(n,k)$ is the power weightings, while $I_{azi}(n,k)$ and $I_{ele}(n,k)$ are the azimuth ad elevation power intensity of each time-frequency instant, respectively. The $W(n,k)$, $C(n,k)$, where $C = X,Y,Z$ are the time-frequency instant in the four channels. The $conj$ represents conjunct, while Re demonstrates real part. The final DOA estimation is based on a weighted average of the values in Eq. (10), where the weights are based on the magnitude of the corresponding time-frequency component in the W channel (this reduces the impact on the DOA estimation of low magnitude components not corresponding to direct sound [15]). The segmental length of direct sound in time domain is 5 ms referenced by [21].

IV. Results

When comparing the MSE for different lengths of DS segments in Table 1, it is clear that the model incorporating MRSTFT and EDT losses achieves the lowest MSE for both 5 ms and 50 ms segment lengths. This model also records the lowest DOA azimuth and elevation errors, as DOA and MSE metrics evaluate the same DS section (5 ms). Furthermore, the model with MRSTFT, RT30, EDT, and DRR losses achieves the minimum errors in RT, EDT, and DRR measurements. The objective results in Table 1 also reveal that the model with MRSTFT and RT30 losses achieves the second-lowest RT30 errors, while the model combining MRSTFT and DRR losses achieves the second-lowest DRR error. Although the model with MRSTFT and EDT does not reach the lowest EDT error, the model combined MRSTFT, RT, EDT, and DRR losses does. Additionally, the model with MRSTFT, RT30, and DRR losses achieves the lowest MRSTFT error. The DROQM scores for all models exceed 0.4, indicating that the quality is acceptable.

Fig. 1 illustrates the mean value of MAE with 95% confidence intervals for the model using MRSTFT as main loss function, with various combinations of acoustic parameters as auxiliary loss functions over time for each channel. Comparing Fig. 1(a), 1(d), 1(g), and 1(j), the model combining MRSTFT, RT30, EDT, and DRR losses (represented by the magenta line) performs relatively stable to the model using only the MRSTFT loss (blue line). In Fig. 1(b), 1(e), 1(h), and 1(k), the model that combines MRSTFT and EDT losses (green line) consistently performs well compared to models that combine MRSTFT with only one other acoustic parameter. This model achieves the lowest mean value of MSE from 0 to 0.05 seconds, covering the DS and some early reflections, outperforming all other models in this time frame. This corresponds to the lowest MSE across different segment lengths and DOA errors. Additionally, when considering the MRSTFT loss combined with two acoustic parameter losses, the model with MRSTFT, EDT, and DRR losses (magenta line) shows greater stability with fewer errors throughout the entire signal duration, as shown in Fig. 1(c), 1(f), 1(i), and 1(l). This stability is more observable compared to the combination of MRSTFT, RT30, and EDT losses (blue line), as well as the combination of MRSTFT, RT30, and DRR losses (black line).

V. Discussion

Different acoustic parameters contribute to improvements in various aspects of the performance of models. Generally, loss combinations that include the acoustic parameter EDT consistently result in lower errors compared to those incorporating RT30 and DRR. When specific evaluations aim to achieve the lowest error values, selecting the corresponding acoustic parameter can reach the goal. A significant challenge in training a network with FOA RIR data is managing multiple channels to maintain the directional relationships among B-format channels. The experiments demonstrate that focusing on generation accuracy for each channel enhances the overall accuracy of generated FOA RIRs.

VI. Conclusions and Future Work

In this work, a novel CGAN-based network is proposed to generate FOA RIRs for specified receiver and source positions, as well as room dimension. The network employs a generator loss that includes adaptive weightings of various combinations of acoustic parameter loss functions, in addition to the CGAN loss. Objective results show that this network facilitates the generation of FOA RIRs with controlled MRSTFT and acoustic parameters through adaptive weightings. Specific acoustic parameters are utilized to manage the corresponding characteristics of the generated FOA RIRs. The models incorporating MRSTFT and EDT, as well as those combining MRSTFT, RT30, EDT, and DRR, outperform other loss combinations. The EDT acoustic parameter demonstrates better stability and fewer errors compared to models using alternative acoustic parameter losses. Future work will investigate alternative loss functions to further improve FOA RIR generation and explore extending this approach to produce higher-order Ambisonics using neural networks.

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] 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, Apr. 1979, doi: 10.1121/1.382599.

[2] J. Zhao, X. Zheng, C. Ritz, and D. Jang, "Interpolating the Directional Room Impulse Response for Dynamic Spatial Audio Reproduction," *Applied Sciences*, vol. 12, no. 4, Art. no. 4, Jan. 2022, doi: 10.3390/app12042061.

[3] A. Southern, J. Wells, and D. Murphy, "Rendering walk-through auralisations using wave-based acoustical models," in *2009 17th European Signal Processing Conference*, IEEE, 2009, pp. 715–719.

[4] J. G. Tylka and E. Y. Choueiri, "Soundfield Navigation using an Array of Higher-Order Ambisonics Microphones," *Los Angeles*, p. 10, 2016.

[5] T. McKenzie, N. Meyer-Kahlen, R. Daugintis, L. McCormack, S. Schlecht, and V. Pulkki, "Perceptually informed interpolation and rendering of spatial room impulse responses for room transitions," presented at the 24th International Congress on Acoustics (ICA), Gyeongju, Korea, Oct. 2022.

[6] 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.

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

[8] A. Ratnarajah, Z. Tang, and D. Manocha, "IR-GAN: Room Impulse Response Generator for Far-Field Speech Recognition," in *Interspeech 2021*, ISCA, Aug. 2021, pp. 286–290. doi: 10.21437/Interspeech.2021-230.

[9] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv.org. Available: https://arxiv.org/abs/1411.1784v1

[10] 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.

[11] 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.

[12] 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.

[13] Y. He, J.-X. Zhong, Z. Dai, N. Trigoni, and A. Markham, "SoundNeRirF: Receiver-to-Receiver Sound Neural Room Impulse Response Field," Sep. 2022.

[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] X. Zheng, C. Ritz, and J. Xi, "Encoding and communicating navigable speech soundfields," *Multimed Tools Appl*, vol. 75, no. 9, pp. 5183–5204, May 2016, doi: 10.1007/s11042-015-2989-3.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Patrick A. Naylor and Nikolay D. Gaubitch, *Speech Dereverberation*. Springer London, 2010.

[21] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2110–2117, Oct. 2002, doi: 10.1121/1.1506692.

[22] 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.

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

[24] 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.

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