Audio Engineering Society

Convention Express Paper 305

Presented at the AES 157th Convention
2024 October 8–10, New York, NY, USA

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

A General Overview of Methods for Generating Room Impulse Responses

Vlad Baran[1,2], Wieslaw Woszczyk[1,2], and Richard King[1,2]

[1]*McGill University*
[2]*Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT)*

Correspondence should be addressed to Vlad Baran (vlad.baran@mail.mcgill.ca)

ABSTRACT

The utilization of room impulse responses has proven valuable for both the acoustic assessment of indoor environments and music production. Various techniques have been devised over time to capture these responses. Although algorithmic solutions have been in existence since the 1960s for generating synthetic reverberation in real time, they continue to be computationally demanding and in general lack the accuracy in comparison to measured authentic Room Impulse Responses (RIR). In recent times, machine learning has found application in diverse fields, including acoustics, leading to the development of techniques for generating RIRs. This paper provides a general overview, of approaches and methods for generating RIRs, categorized into algorithmic and machine learning techniques, with a particular emphasis on the latter. Discussion covers the acoustical attributes of rooms relevant to perceptual testing and methodologies for comparing RIRs. The paper is designed to offer a general overview for those interested in RIR generation for music production purposes, with future work considerations also explored.

1 Introduction

Accurate estimation of impulse responses (IRs) is important for various audio and speech signal processing applications, including modeling scene and room acoustics, localizing sound sources, and achieving realistic audio spatialization. Typically, these IRs are spatio-temporal, meaning that each spatial location in a 3D environment corresponds to a short time-domain signal. However, measuring IRs for every spatial point in a 3D scene is often impractical due to setup complexity and cost, especially considering that the setups themselves are not always easily portable. As a result, researchers have turned to computational models for reliable IR estimation and prediction. As a result, there has been a surge in the development of room impulse response (RIR) generators, which create realistic RIRs for specific acoustic environments.

In this paper, two main approaches for estimating RIRs are explored: geometrical acoustic modeling and ma-

chine learning. Over the past few years, machine learning has gained popularity due to increased computational power and the availability of large databases. This paper highlights some approaches while describing the methods used.

2 Scope of the review

The purpose of the overview is to compare fundamental models of geometrical room acoustic modeling techniques with emerging machine learning approaches. Emphasis is placed on the newly emerging techniques that leverage machine learning to generate room impulse responses. Time-wise the span of this paper's focus is after the invention of the Generative Adversarial Model(GAN) [1] in 2014 until the beginning of 2024. To provide context for these novel approaches, geometrical acoustic modeling is introduced. The paper covers theoretical aspects of both geometrical modeling and machine learning, with a focus on algorithm performance. This paper will inform the reader on the accessibility of the new emerging algorithms as well as the multiple models that can be implemented.

3 Geometrical Acoustics modeling techniques

Room impulse responses find diverse applications, ranging from architectural acoustics measurements to capturing room properties for music production. The attributes that can be derived from the impulse responses are,upon others, reverberation time (T60) and clarity (C80). In musical contexts, auralization involves convolving the recorded signal (often without natural reverberation) with the RIR, creating a simulated performance within the captured acoustic environment [2]. As defined by Kleiner Mendel, "auralization is the process of rendering audible, by physical or mathematical modelling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space" [3]. The fundamental principles of auralization systems were initially introduced by Schroeder et. al. in 1962 [4] and revisited by Schroeder in 1970 [5]. Sabine, in the early 1900s, employed ray-based acoustics and Schlieren photography [6] [7] to explore sound propagation paths. His visualizations depicted wavefronts augmented with rays perpendicular to them, akin to an early form of computerized ray tracing. The concept of a mean free path, measuring the average distance a ray travels between successive reflections in a given geometry, is closely tied to geometrical acoustics (GA). This measure underlies the early reverberation time formulas developed by Sabine and Eyring [8]. The validity of the geometrical acoustic approach, particularly when sound wavelengths are small compared to room dimensions, was demonstrated in 1929 [9]. Allred and Newhouse presented the first computer model for determining mean free path lengths in 1958 [10].

3.1 Image Source Method (IS)

This method was developed in a study for electrostatics and early acoustics work presented by Carsalw in 1899 [11]. Cremer [12] presented the calculation of sound pressure as function of time in a rectangular "shoe box" room in 1948. The initial method was developed to predict the sound pressure levels caused by non directional source of known sound power radiation between absorbing parallel planes. This solution is exact for spherical waves interacting with a rigid surface. The image source solution is also asymptotically correct for high frequencies. The image-source method is the process of applying recursively the image source solution. The method reflects the sound source against all surfaces in the model, this results in a set of image sources which are again reflected against the same surfaces. This process repeated until a condition is met, such as response length or reflection order. The result of the image source method can be viewed as a hierarchical tree with the sound source as a root and each branch representing an IS. Computing IRs for polygonal structures using this method was done firstly by Borish in 1984 [13]. The recursive method is applied for the rectangular spaces but there are additional steps that need to be completed. One important consideration is that in irregular shaped rooms there may be parts that are not "visible" from the sound source position [2].

3.2 Ray Tracing Techniques

In contrast with image source method that provides a geometrical solution that describes all specular reflection paths, the ray tracing methods are stochastic and perform a random sampling of all possible reflection paths. The first appearance of this method was in 1958 by Allred and Newhouse [10], however the full description on obtaining time-energy responses in 3D space was done by Krokstad et al in 1968 [14].

The main principle of this method is to case rays from the sound source reflect them and store the paths. The ray directivity pattern can be pre-determined or random. The rays continue after the first reflection and interactions between them and detectors are computed to track which reflections arrive to the receiver. Termination of rays can be done using two methods [15]. The first method attributes energy properties to each rays and depending on what type of material it reflects off that energy decreases until it reached a preset threshold which terminates the ray. Frequency bands can be implemented so the reflections can affect bands differently. The second method attributes probability of annihilation to each ray and depending on the material it reflects off it may be completely absorbed.

One important advantage over the image source method is that arbitrary reflections can be simulated, for example diffraction. There are a number of methods for calculating the diffuse reflections. The basic ray tracing technique encounters challenges related to volumetric detectors. In an ideal scenario, detectors would be point-like, but due to the small probability of a ray intersecting a point, detectors are designed with a volume. However, this approach introduces issues, as volumetric detectors may register paths that are invalid for the listener's specific position [16] [17]. Another issue is computational power, when calculating all diffuse reflections the number of rays increases exponentially and the required calculations increase accordingly [18]. An extensive comparison has been done by Stephenson between image source and ray tracing techniques concluding that the latter presents a superior performance in most cases [19].

3.3 Beam Tracing

There are two main techniques for beam tracing, one closer to image source method and another to ray tracing [2]. In the image source technique discrete rays are used to plot the path of reflections, it is obvious that some paths will be missed since the number of rays is limited. Beam tracing expands these rays into volumetric objects that are captured by a point receiver. First published in 1973 by Haviland and Thanedar [20], they suggest using conical shaped beams instead of rays. One advantage of this approach is that surfaces like spheres can be easily computed with the entire space around the object being covered in a volumetric beam. A common problem that appears when using this technique is that the energy at the receiver is overestimated due to the fact that volumetric beams do not split realistically when hitting multiple reflective surfaces [21] [22]. This was solved in 2000 by two algorithms the adaptive beam tracing by Drumm and Lam [23] and adaptive pyramid tracing by Campo et al [24]. Both techniques are splitting the original beams into smaller beams upon reflection. This means the algorithms to be computationally ineffective since the use of volumetric beams can increase very rapidly. A solution to this is proposed by Chandak et al. [25] where a two stage process is implemented, in the first phase a volumetric beam is used and after specular reflections are calculated the image sources are verified as in the original image source algorithm.

Beam tracing can be used as an optimisation stage for the image source method. The purpose of this stage is to reduce the number of image sources and thus reduce the computation load. The reduction is done for IS that do not have valid reflection paths. The process is similar to the image source technique, the first order image sources are created and after that volumetric beams are created for each IS by the edges of the surfaces that created the first order IS. The beams reflect only on surfaces that are inside the beam and not all surfaces as in the original image source technique. This reduces the computational strain and narrows the image source tree with more efficient reflections in more complex scenarios. This technique was presented in 1996 by Stephenson [21]. The technique called Quantized Pyramidal Beam Tracing addresses also diffraction modeling by introducing new beams at diffracting edges of the room geometry. In order to avoid a large number of beams the concept of quantization and reunification of beams is implemented, diffracted and reflected beams are spatially unified. The unification is applied only at higher order reflections and such it does not affect the accuracy of the model. Funkhouser et al. presented a working model of this technique in 1998 [26] [27]. This auralization system was designed to consider specular and edge diffraction.

3.4 Radiosity Technique

Originally developed for computer graphics to simulate global illumination [28] [29], this approach is in contrast to the image source method since it is surface based and assumes purely diffuse reflections. The first study to analyze the radiosity technique was the doctoral thesis of Moore in 1984 [30]. The thesis provides a very detailed description of the implementation

by first defining a new Pascal-like programming language called Sabine, in which the modeling tool is coded. Another significant advancement in this technique was made by Nosal et al. in 2004, detailing how it predicts sound fields in non-rectangular rooms with ideally diffusely reflective boundaries [31]. The radiosity technique possesses this feature: sound energy radiating from a surface is angle-independent, as implied by its basic assumption. This property significantly reduces memory requirements, making the technique well-suited for real-time applications. In such scenarios, radiosity simulations can be conducted as a pre-processing step, allowing sound energy at the listener's position to be gathered at interactive rates, even when the listener is in motion.

3.5 Acoustic Radiance Transfer

The basic principles of acoustic radiance transfer (ART) are similar to the radiosity technique. The concept of arbitrary reflections properties is introduced in this method, which makes it more practical for real life applications in contrast to radiosity technique. Proposed by Siltanen et al. the basic principles of ART remain the same as in the radiosity, and the resulting reflections are now a function of the outgoing direction [32]. However this increases the computational load since the distribution of reflections is more complex. Another difference is the behaviours of surface reflections, emitted energy from a point depends on the incident angle of the original wave. For those reasons simulations are more efficient in frequency domain rather than time domain since the algorithm takes into account reflection properties of surfaces [33].

4 Machine Learning Modeling techniques

In practical terms, synthetic RIR generators excel at capturing sound propagation phenomena in regularly shaped or largely empty rooms [2]. Creating realistic sound reflections in complicated environments, like staircases, is more difficult because it requires more accessible and less computationally complex simulations. Consequently, there is a need for alternative synthetic RIR generator methods capable of accurately modeling sound effects in complex environments.

4.1 Models

Machine learning has gained prominence, especially with large language models, leading to the increasing popularity of generative artificial intelligence. A neural network consists of multiple layers of interconnected neurons[34]. The initial layer is known as the input layer, while the last layer is the output layer. Intermediate layers, sandwiched between the input and output layers, are referred to as hidden layers. Networks with multiple layers of neurons between the input and output are considered to be a 'deep neural networks'. Each neuron has an adjustable bias parameter that learns during the training process. Neurons communicate through weighted connections, which are also fine-tuned as data is fed into the network. Activation functions, typically nonlinear, play a crucial role in extracting complex patterns from the data. Lastly, the cost function evaluates a model's performance by measuring the difference between predicted and actual values [34].

The multi-layer perceptron(MLP) network is very similar to the general architecture described earlier. It contains layers of neurons that are connected by weights. There is no feedback loop, the output of a neuron does not affect the neuron itself, this is called a feed-forward network. The major use cases of MLP are pattern classification, recognition, prediction and approximation [35]. Convolutional Neural Netowkrs(CNN) are neural networks inspired by visual perception [36] [37]. The architecture is somehow similar to the MLP but there are 3 key differences: 1. local connections - Neurons are no longer connected to all neurons of the previous layer, but only to a small number of neurons, which is effective in reducing parameters and speed up convergence. 2. Weight sharing—a group of connections can share the same weights, which reduces complexity of the network further. 3. Down-sampling dimension reduction a pooling layer down-samples an image, which can reduce the amount of data while retaining useful information [38]. Lastly the Generative Adversarial Network(GAN) is a more complex network that can contain other neural networks inside of it [1]. GANs often comprise a generator and a discriminator that learn simultaneously. The generator tries to capture the potential distribution of real samples, and generates new data samples. The discriminator is often a binary classifier, discriminating real samples from the generated ones. Besides image generation [39] GANs have been used in speech and language processing [40] [41],

malware detection [42], and chess game programming [43].

4.1.1 Acoustical Parameter Estimation Using Deep Neural Networks

The work of Yu and Kleijn shows an algorithm that can estimate the acoustical properties of a room in the aid of generating room impulse responses [44]. This approach has two stages that contain different neural networks. Firstly, the approximation of room geometry using convolutional neural networks(CNN). Because the real shape of a room is more complex than a shoe-box due to the obstacles and various absorption coefficients the authors decided to treat the relationship between room geometry and RIRs as probabilistic. When considering the effect of room geometry on impulse responses each room creates a different set of time arrivals at the listener position. Since CNNs are optimised to detect patterns in data the authors have chose that architecture for room geometry estimation from RIRs. In order to address the reflection coefficients of different walls in a room a RIR is divided into four frequency bands. The authors treat the reflection coefficients as features that can be extracted directly from the file by multi-layered perceptrons(MLP).

For training the authors choose the I. Szoke et al. database [45]. Since the focus is on medium to small room RIR generation the two large rooms from that database were excluded. The six remaining rooms contain an average of 155 RIRs with 31 receiver positions with the longest RT30 being 1.8 s. This amount of data is insufficient to properly train a machine learning algorithm[46], this is a common problem in machine learning applications. The authors also made an extra 400,000 using the image source method. As validation multiple experiments were run to test the performance of this approach. The algorithm was compared to the image source method approach in [47] done with [48]. The experiment evaluated average run time, system requirements and estimation error. Along 600 experiments it was observed that the accuracy of RIR estimation was approximately the same while the runtime was 1,000 times smaller for the proposed method. The approximation of room geometry from real-world RIRs was also tested. Using the BUT ReverbDB database, 6 rooms that contained multiple RIRs. The root mean square error was for single dimension was minimum 0.05 m and the maximum 0.26 m. Based on the experimental results, the proposed method achieves nearly identical accuracy while demanding approximately 10,000 times less computational effort after training.

4.1.2 IR-GAN and FAST-RIR

Ratnarajah et al. peresent a GAN-based RIR generator that can simulate real and existing spaces [49]. The architecture is based on WaveGAN and it is evaluated for automatic speech recognition [50]. Trained by real world RIRs gathered from the BUT ReverbDB [45] and AIR dataset [51]. In order to limit the amount of impulses that cannot exist in reality, the network is constrained by the following relevant acoustical parameters that can be extracted from a RIR:

- Early-decay-time (EDT)
- Early-to-late index (CTE)
- Reverberation time (T60)
- Direct-to-reverberant ratio (DRR)

The data that is used for training is sampled at 16Khz with audio samples fed into the GAN in the form of 32 bit float point vector of length 16384. This allows the model to create RIRs that are slightly longer than one second. The training data is composed of 1209 RIRs from BUT ReverbDB and 68 RIRs from AIR, all of those impulse responses characterize small to medium seized rooms. All the training and testing was done on 32 Intel(R) Xeon(R) Silver 4208 CPUs @ 2.10 GHz and 2 GeForce RTX 2080 Ti GPUs.

The performance test is done using modified Kaldi LibriSpeech ASR recipe that tests the automatic speech recognition in reverberant spaces [52]. The results are based on Word Error Rate(WER) [53] and are compared to geometrical algorithms that were instructed to simulate the same acoustical environments. The authors report a 8.95% performance improvement in comparison with the geometrical acoustic simulator, performance further improves when the RIRs are combined arriving to 14.3%.

Another model created by the same authors is named FAST-RIR [46]. This approach can generate specular and diffuse reflections with arbitrary T60 by having the dimensions of the room. Similar to the previous model it is focused on small-medium rooms, with the

evaluation being done by automatic speech detectors Google Speech API and Microsoft Speech API. The inputs are room geometry, listener and speaker positions, and T60 with a RIR as output, the architecture is based on StackGAN, an image generation machine learning algorithm [54]. The generation of impulses was done on Intel(R) Xenon(R) CPU E52699 v4 @ 2.20 GHz and a GeForce RTX 2080 Ti GPU. Evaluation is done in two ways, calculating the T60 error and calculating the WER from automatic speech recognition. The T60 error is calculated using the ISO 3382-1:2009 and is the comparison between the value given to the algorithm and the resulting value of the generated RIR. Below 0.25 s the error is high (0.068s) when compared to the error for values higher than 0.25 s(0.021s). The ASR evaluation is done using convolution between the LibriSpeech database examples and different RIRs. The other algorithms tested are image source method [47], difuse acoustic simulator [55], and gpuRIR [56]. The authors report that the performance of FAST-RIR is similar to state of the art diffuse acoustic simulator regarding WER.

4.1.3 Image2Reverb

Singh et. al. have developed an machine learning algorithm that can generate RIRs directly from images of acoustic spaces [57]. The model is based on a conditional GAN that takes as input an image and outputs a spectrogram. The architecture that contains an encoder, generator, and conditional discriminator is similar to Mentzer et al. which is designed to compress images [58]. The generator and discriminator are designed after the GANSynth model [59]. Audio is represented as a log magnitude spectrogram. Firstly the audio is sampled at 22.050kHz and truncated to 5.94 s then a Fourier transform with window size 1024 and hop size 256 resulting in 512x512 spectrograms. Finally a log(|X|) created the magnitude spectrogram. The encoder module combines image feature extraction with depth estimation to produce latent vectors from two dimensional images of scenes. Training the network is done on 8 NVIDIA 1080 Ti GPUs. The depth estimation is done using another pre-trained neural network called Monodepth2 which created a one-channel depth map corresponding to the given image [60]. The main feature extractor that is responsible for scene recognition and categorisation is ResNet50 [61] which was pretrained on Places365 [62]. This is a image recognition model that was trained on a a database of over 7 million images. The database was composed of 265 different spaces totalling 1169 images and 738 RIRs. This resulted in 11234 RIR to image paired examples structured into train-validation-test split groups of 9743-154-1957. Examples came from the OpenAIR dataset and other sources [63].

In this paper a perceptual test was conducted designed to estimate comparative perceptual plausibility between the generated RIRs and the measured ones. A panel of 31 expert listeners was asked to rate 16 examples of an anechoic source convcolved with the output IR, as well as the ground truth IR. The rating was done on a scale from 1 to 5 based on 1) reverberation quality, and 2) realism or “match” between their expected reverb based on the image and the presented signal with reverb applied. Results show a preference of the listeners towards the real RIRs in all four examples "large", "medium", "outdoor", "small" with one exception where the results are split. Since the authors describe the anechoic stimulus that was convolved as "arbitrary" there is little information about the type of audio that was used as well as the listening environment.

4.1.4 Deep Impulse Responses

This paper analyses the process of interpolation between positions of meausred IRs [64]. While not generating RIRs from geometric dimensions this paper presents an important approach on expanding on existing RIRs. The authors also describe parametrisation of IRs for compact storage and the generation of IRs in noisy environments, those aspects are outside the scope of this paper. A similar approach is implemented by Sanaguano-Moreno et al. also using the head related impulse responses [65].

The data contains 9,720 head related impulse responses(HRIR) that are captured by the team in an anechoic chamber. The impulse responses are measured on a sphere around a listener, each of which is a two-channel FIR with 400 taps per channel. All audio data was sampled at 48KHz. The model is a MLP that consists of six fully connected layers with 512 hidden units each. As input the networks receives 3D spatial coordinates x,y,z and estimates the the t-th sample of the finite impulse response. The evaluation is done using Signal-to-Distortion Ratio [66]. From the 9,720 measured IRs the authors train the network using data generated from a limited number of the measured IRs

and evaluate how accurately the network can predict the impulse responses at positions that were not part of the training data. The model is compared to neural renderer in [67] and bilinear interpolation [68]. The results show that the method proposed outperforms the neural renderer in [67] and, with sufficient training examples, outperforms biliniar interpolation.

4.1.5 Room impulse response reconstruction with physics-informed deep learning

Physics-informed neural networks are trained to solve partial differential equations [69]. In the context of acoustics, these networks are specifically trained to solve the wave equation within the appropriate environment [70]. By incorporating physical principles during model training, these networks have a better chance of accurately reconstructing sound fields. As a result, they are promising candidates for sound field reconstruction, potentially offering improved accuracy and computational efficiency compared to traditional methods. Additionally, since they function as approximators [71], they allow for a continuous representation of the sound field they aim to reconstruct (grid-free mapping). The authors of this paper utilize this type of machine learning model to reconstruct sound fields from any position within a room in real time.

Traditionally, a Physics-Informed Neural Network (PINN) is often modeled as a multi-layer perceptron (MLP). However, MLPs can suffer from stiffness in gradient flow dynamics. To address this limitation, the authors propose a specialized neural network architecture known as the modified multi-layer perceptron (mMLP) introduced by Wang et al [72], which draws inspiration from transformer attention mechanisms [73]. This modification aims to enhance the effectiveness of PINNs.

The experimental dataset used for training and evaluating the PINN comprises of RIRs measured in a dampened room at the Technical University of Denmark. These RIRs were acquired using a Universal Robots UR5 robotic arm equipped with a 1/2" omnidirectional Brueel & Kjaer pressure microphone, covering a grid of 900 positions within the room. However, only 100 of these positions were utilized as training data for the network.

Regarding the simulated data, the network architectures employ 800 neurons in each layer, except for the input layer (with 3 neurons) and the output layer (with 1 neuron). The training process involves 150,000 iterations, executed on a single 32 GB NVIDIA V100 GPU. For the simulated data, it takes approximately 9 hours, while for the experimental data, it takes around 6 hours, with memory utilization capped at 25% of the total capacity.

In this study, the authors compare the presented model with another machine learning approach called Deep Image Prior [74]. Deep Prior is based on image inpainting, treating the measured RIRs as an incomplete image represented by a binary matrix that activates or deactivates pixels based on data availability. Notably, instead of using a linear array of measured RIRs for Deep Prior fitting the authors opt for multiple 2D pressure snapshots to compare the two approaches. No subjective tests were conducted in this comparison.

The objective comparison results demonstrate that Physics-Informed Neural Networks can effectively characterize sound fields, reconstructing essential time-domain field quantities such as particle velocity and instantaneous intensity. Notably, PINNs exhibit high accuracy in reconstructing the early portion of room impulse responses, outperforming other data-driven priors in the form on neural networks. In summary, this study underscores the reliability and versatility of PINNs as a valuable tool for sound field reconstruction tasks, applicable in both simulated and real-world acoustic environments.

5 Data

Developers of neural networks often grapple with the challenge that training data alone is insufficient to achieve the desired results. In this paper, we explore various methods for creating impulse responses, each accompanied by its unique approach to building a database of RIRs. The demand for larger and more comprehensive RIR databases continues to grow [44] [46].

As illustrated in Table 1, the number of RIRs needed to train a machine learning network varies based on the specific model and desired outcomes. In the Model column the papers presented in this overview are referenced by abbreviation: Acoustical Parameter Estimation(APE), PINN(Physics-informed neural networks), DIR(Deep Impulse Responses).

In this study, various approaches have either built their own custom databases of RIRs or leveraged publicly

Model	Real #RIRs	Generated #RIRs	#Environments	Train-Development-Test
APE	155	400,000	6	280,000-80,000-40,000
IR-GAN	1269	0	N/A	773-194-242
FAST-RIR	0	75,000	15	30,000-30,000-15,000
Image2Reverb	11,234	0	265	9743-154-1957
PINN	900	0	1	800-0-100
DIR	9,720	0	1	6000-N/A-N/A

Table 1: Training data for RIR Generators. Train-Development-Test indicates the distribution of data as presented to the network.

Name	Real #RIRs	#Rooms	T60(s)	IR Type
ACE	700	7	0.33–1.25	ESS
AIR	214	6	0.12–0.78	MLS
REVERB	24	3	0.25–0.70	N/A
RWCP	3000	9	0.01–2.00	TSP
BUT ReverbDB	1300	8	0.59–1.85	ESS
TUD	900	1	0.52	ESS

Table 2: Publicly accessible RIR databases.

available ones. Table 2 showcases publicly accessible RIR databases, including those documented by the BUT ReverbDb paper[45], along with the recent addition of the TUD(Technical University of Denmark) database (released in early 2024)[70]. While this compilation is not exhaustive, it serves as a foundational resource for researchers and practitioners developing working models based on RIRs.

6 Conclusions and Discussion

Is this paper multiple methods of creating room impulse responses were presented, with the focus being on machine learning algorithms. As it can be observed from the papers presented on chapter 5 the majority of authors measure the performance of the RIR generation using mean average error or some similar objective measurement between the geometrical acoustics simulator and the machine learning algorithm. There are few perceptual tests being done and from the few that are listed in this analysis the machine learning approach performs poorly. For music production, perceptual tests play a crucial role because objective measurements alone are inadequate for assessing the realism of a generated room impulse response. Simultaneously, the adoption of machine learning-based RIRs in music production remains limited, resulting in a reduced need for perceptual testing. Currently, there appears to be a gap between the communities that investigate accurate room acoustic modeling techniques and those who focus on music production. I believe that these areas would benefit from collaboration.

To close the gap, we need further investigation to evaluate the quality and authenticity of these generated RIRs. Additionally, exploring models capable of training on existing data to produce realistic results would enhance the flexibility and versatility of convolution-based reverberators.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," *Advances in neural information processing systems*, 27, 2014.

[2] Savioja, L. and Svensson, U. P., "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, 138(2), pp. 708–730, 2015, ISSN 0001-4966, 1520-8524.

[3] Mendel, K., Bengtinge, D., and Peter, S., "Auralization-An overview," *Journal of the Audio Engineering Society*, 41, pp. 861–875, 1993.

[4] Schroeder, M., Atal, B. S., and Bird, C., "Computers in room acoustics," in *Proceedings of the 4th International Conference on Acoustics*, 1962.

[5] Schroeder, M. R., "Digital simulation of sound transmission in reverberant spaces," *The Journal of the acoustical society of america*, 47(2A), pp. 424–431, 1970.

[6] Sabine, W. C., *Collected papers on acoustics*, Harvard university press, 1922.

[7] Davis, A. and Fleming, N., "Sound pulse photography as applied to the study of architectural acoustics," *Journal of Scientific Instruments*, 3(12), p. 393, 1926.

[8] Knudsen, V. O., "Recent developments in architectural acoustics," *Reviews of Modern Physics*, 6(1), p. 1, 1934.

[9] Eyring, C. F., "Reverberation time in "dead" rooms," *The Journal of the Acoustical Society of America*, 1(2A), pp. 217–241, 1930.

[10] Allred, J. C. and Newhouse, A., "Applications of the Monte Carlo method to architectural acoustics. II," *The Journal of the Acoustical Society of America*, 30(10), pp. 903–904, 1958.

[11] Cakslaw, H., "Some Multiform Solutions of the Partial Differential Equations of Physical Mathematics and their Applications," in *Proceedings of the London Mathematical Society*, volume 1, pp. 121–165, Oxford University Press, 1898.

[12] Cremer, L. and Müller, H. A., "Die wissenschaftlichen grundlagen der raumakustik," *Acta Acustica united with Acustica*, 1948.

[13] Borish, J., "Extension of the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, 75(6), pp. 1827–1836, 1984.

[14] Krokstad, A., Strom, S., and Sørsdal, S., "Calculating the acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, 8(1), pp. 118–125, 1968.

[15] Vorländer, v. M., "Die Genauigkeit von Berechnungen mit dem raumakustischen Schallteilchenmodell und ihre Abhängigkeit von der Rechenzeit," *Acta Acustica united with Acustica*, 66(2), pp. 90–96, 1988.

[16] Vorländer, M., "Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm," *The Journal of the Acoustical Society of America*, 86(1), pp. 172–178, 1989.

[17] Lehnert, H., "Systematic errors of the ray-tracing algorithm," *Applied Acoustics*, 38(2-4), pp. 207–221, 1993.

[18] Kuttruff, H., "Simulierte Nachschallkurven in Rechteckraeumen. mit diffusem Schallfeld," *Acustica*, 25, pp. 333–342, 1971.

[19] Stephenson, U., "Comparison of the mirror image source method and the sound particle simulation method," *Applied Acoustics*, 29(1), pp. 35–72, 1990.

[20] Haviland, J. K. and Thanedar, B. D., "Monte Carlo applications to acoustical field solutions," *The Journal of the Acoustical Society of America*, 54(6), pp. 1442–1448, 1973.

[21] Stephenson, U. M., "Quantized pyramidal beam tracing-a new algorithm for room acoustics and noise immission prognosis," *Acta Acustica united with Acustica*, 82(3), pp. 517–525, 1996.

[22] Dalenbäck, B.-I. L., "Room acoustic prediction based on a unified treatment of diffuse and specular reflection," *The Journal of the Acoustical Society of America*, 100(2), pp. 899–909, 1996.

[23] Drumm, I. A. and Lam, Y. W., "The adaptive beam-tracing algorithm," *The Journal of the Acoustical Society of America*, 107(3), pp. 1405–1412, 2000.

[24] Campo, N., Rissone, P., and Toderi, M., "Adaptive pyramid tracing: a new technique for room acoustics," *Applied Acoustics*, 61(2), pp. 199–221, 2000.

[25] Chandak, A., Antani, L., Taylor, M., and Manocha, D., "FastV: From-point Visibility Culling on Complex Models," in *Computer Graphics Forum*, volume 28, pp. 1237–1246, Wiley Online Library, 2009.

[26] Funkhouser, T., Carlbom, I., Elko, G., Pingali, G., Sondhi, M., and West, J., "A beam tracing approach to acoustic modeling for interactive virtual environments," in *Proceedings of the*

25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '98, p. 21–32, Association for Computing Machinery, New York, NY, USA, 1998, ISBN 0897919998, doi:10.1145/280814.280818.

[27] Funkhouser, T., Tsingos, N., Carlbom, I., Elko, G., Sondhi, M., West, J. E., Pingali, G., Min, P., and Ngan, A., "A beam tracing method for interactive architectural acoustics," *The Journal of the Acoustical Society of America*, 115(2), pp. 739–756, 2004.

[28] Goral, C. M., Torrance, K. E., Greenberg, D. P., and Battaile, B., "Modeling the interaction of light between diffuse surfaces," *ACM SIGGRAPH computer graphics*, 18(3), pp. 213–222, 1984.

[29] Immel, D. S., Cohen, M. F., and Greenberg, D. P., "A radiosity method for non-diffuse environments," *Acm Siggraph Computer Graphics*, 20(4), pp. 133–142, 1986.

[30] Moore, G. R., *Approach to the analysis of sound in auditoria: model design and computer implementation.*, Ph.D. thesis, 1985.

[31] Nosal, E.-M., Hodgson, M., and Ashdown, I., "Improved algorithms and methods for room sound-field prediction by acoustical radiosity in arbitrary polyhedral rooms," *The Journal of the Acoustical Society of America*, 116(2), pp. 970–980, 2004.

[32] Siltanen, S., Lokki, T., Kiminki, S., and Savioja, L., "The room acoustic rendering equation," *The Journal of the Acoustical Society of America*, 122(3), pp. 1624–1635, 2007.

[33] Siltanen, S., Lokki, T., and Savioja, L., "Frequency domain acoustic radiance transfer for real-time auralization," *Acta Acustica united with Acustica*, 95(1), pp. 106–117, 2009.

[34] Zhou, Z.-H., *Machine learning*, Springer nature, 2021.

[35] Ramchoun, H., Ghanou, Y., Ettaouil, M., and Janati Idrissi, M. A., "Multilayer perceptron: Architecture optimization and training," 2016.

[36] Celeghin, A., Borriero, A., Orsenigo, D., Diano, M., Méndez Guerrero, C. A., Perotti, A., Petri, G., and Tamietto, M., "Convolutional neural networks for vision neuroscience: Significance, developments, and outstanding issues," *Frontiers in Computational Neuroscience*, 17, p. 1153572, 2023.

[37] Matsugu, M., Mori, K., Mitari, Y., and Kaneda, Y., "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural networks*, 16(5-6), pp. 555–559, 2003.

[38] Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J., "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, 33(12), pp. 6999–7019, 2021.

[39] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

[40] Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D., "Adversarial learning for neural dialogue generation," *arXiv preprint arXiv:1701.06547*, 2017.

[41] Pascual, S., Bonafonte, A., and Serra, J., "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[42] Hu, W. and Tan, Y., "Generating adversarial malware examples for black-box attacks based on GAN," in *International Conference on Data Mining and Big Data*, pp. 409–423, Springer, 2022.

[43] Chidambaram, M. and Qi, Y., "Style transfer generative adversarial networks: Learning to play chess differently," *arXiv preprint arXiv:1702.06762*, 2017.

[44] Yu, W. and Kleijn, W. B., "Room Acoustical Parameter Estimation From Room Impulse Responses Using Deep Neural Networks," 29, pp. 436–447, 2021, ISSN 2329-9290, 2329-9304, doi: 10.1109/TASLP.2020.3043115.

[45] Szoke, I., Skacel, M., Mosner, L., Paliesek, J., and Cernocky, J., "Building and evaluation of a real room impulse response dataset," 13(4), pp. 863–876, 2021, ISSN 1932-4553, 1941-0484, doi: 10.1109/JSTSP.2019.2917582.

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

[47] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, 65(4), pp. 943–950, 1979.

[48] Habets, E., "RIR Generator," `https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator`, 2014, accessed: 2024-06-01.

[49] Ratnarajah, A., Tang, Z., and Manocha, D., "IR-GAN: Room impulse response generator for far-field speech recognition," *arXiv preprint arXiv:2010.13219*, 2020.

[50] Pena, B. and Huang, L., "Wave-GAN: a deep learning approach for the prediction of nonlinear regular wave loads and run-up on a fixed cylinder," *Coastal Engineering*, 167, p. 103902, 2021.

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

[52] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S., "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210, IEEE, 2015.

[53] Klakow, D. and Peters, J., "Testing the correlation of word error rate and perplexity," *Speech Communication*, 38(1-2), pp. 19–28, 2002.

[54] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N., "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

[55] Tang, Z., Chen, L., Wu, B., Yu, D., and Manocha, D., "Improving reverberant speech training using diffuse acoustic simulation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6969–6973, IEEE, 2020.

[56] Diaz-Guerra Aparicio, D., Miguel, A., and Beltran, J. R., "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, 80(4), pp. 5653–5671, 2021, doi:10.48550/arXiv.1810.11359.

[57] Singh, N., Mentch, J., Ng, J., Beveridge, M., and Drori, I., "Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 286–295, IEEE, 2021, ISBN 978-1-66542-812-5, doi:10.1109/ICCV48922.2021.00035.

[58] Mentzer, F., Toderici, G. D., Tschannen, M., and Agustsson, E., "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, 33, pp. 11913–11924, 2020.

[59] Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A., "Gansynth: Adversarial neural audio synthesis," *arXiv preprint arXiv:1902.08710*, 2019.

[60] Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J., "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

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

[62] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A., "Learning deep features for scene recognition using places database," *Advances in neural information processing systems*, 27, 2014.

[63] Murphy, D. T. and Shelley, S., "Openair: An interactive auralization web resource and database," in *129th Audio Engineering Society Convention 2010*, pp. 1270–1278, Audio Engineering Society, 2010.

[64] Richard, A., Dodds, P., and Ithapu, V. K., "Deep Impulse Responses: Estimating and Parameterizing Filters with Deep Networks," (arXiv:2202.03416), 2022.

[65] Sanaguano-Moreno, D., Lucio-Naranjo, J., Tenenbaum, R., and Sampaio-Regattieri, G., "Real-time impulse response: a methodology based on Machine Learning approaches for a rapid impulse response generation for real-time Acoustic Virtual Reality systems," *Intelligent Systems with Applications*, 21, p. 200306, 2024, ISSN 26673053, doi:10.1016/j.iswa.2023.200306.

[66] Yamamoto, K., Irino, T., Matsui, T., Araki, S., Kinoshita, K., and Nakatani, T., "Predicting Speech Intelligibility Using a Gammachirp Envelope Distortion Index Based on the Signal-to-Distortion Ratio." in *INTERSPEECH*, pp. 2949–2953, 2017.

[67] Richard, A., Markovic, D., Gebru, I. D., Krenn, S., Butler, G. A., Torre, F., and Sheikh, Y., "Neural synthesis of binaural speech from mono audio," in *International Conference on Learning Representations*, 2020.

[68] Antonello, N., De Sena, E., Moonen, M., Naylor, P. A., and Van Waterschoot, T., "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10), pp. 1929–1941, 2017.

[69] Borrel-Jensen, N., Engsig-Karup, A. P., and Jeong, C.-H., "Physics-informed neural networks for one-dimensional sound field predictions with parameterized sources and impedance boundaries," *JASA Express Letters*, 1(12), p. 122402, 2021, ISSN 2691-1191, doi:10.1121/10.0009057.

[70] Karakonstantis, X., Caviedes-Nozal, D., Richard, A., and Fernandez-Grande, E., "Room impulse response reconstruction with physics-informed deep learning," *The Journal of the Acoustical Society of America*, 155(2), pp. 1048–1059, 2024.

[71] Hornik, K., Stinchcombe, M., and White, H., "Multilayer feedforward networks are universal approximators," *Neural networks*, 2(5), pp. 359–366, 1989.

[72] Wang, S., Teng, Y., and Perdikaris, P., "Understanding and mitigating gradient flow pathologies in physics-informed neural networks," *SIAM Journal on Scientific Computing*, 43(5), pp. A3055–A3081, 2021.

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

[74] Pezzoli, M., Perini, D., Bernardini, A., Borra, F., Antonacci, F., and Sarti, A., "Deep prior approach for room impulse response reconstruction," *Sensors*, 22(7), p. 2710, 2022.