TECNIACUSTICA ELCHE · 2022

53º CONGRESO ESPAÑOL DE ACÚSTICA
XII CONGRESO IBÉRICO DE ACÚSTICA

SYNTHESIS OF ROOM IMPULSE RESPONSES BY MEANS OF DEEP LEARNING

PACS: 43.55.Br, 43.55.Ka.

Martín Salinas, Ignacio

Dept. Tecnología Electrónica, Universidad Carlos III de Madrid, Av. Universidad, 30, 28911 Leganés (Madrid), España, 100383261@alumnos.uc3m.es

Belloch Rodríguez, Jose Antonio

Dept. Tecnología Electrónica, Universidad Carlos III de Madrid, Av. Universidad, 30, 28911 Leganés (Madrid), España, jbelloc@ing.uc3m.es

Piñero Sipán, Gema

iTEAM, Universitat Politècnica de València, Camí de Vera, s/n, 46022 València, España, gpinyero@iteam.upv.es

Keywords: deep learning, room impulse response, room acoustics.

ABSTRACT

The irruption of deep learning (DL) in the recent years is starting to create a turning point in many areas of digital signal processing. Regarding audio signal processing, machine learning (ML) models in general, and deep neural networks (DNNs) in particular, have shown their potential in the classification of acoustic events, enhancement of speech signals in the presence of noise and acoustic echo cancellation. In this work, we propose a DL model to synthesize room impulse responses (RIRs) of real enclosures by using as input parameters to the network: the Short Time Fourier Transform (STFT) of a measured RIR of the same room, the locations in Cartesian coordinates of the speaker and listener for both input and inferred RIRs, and the room dimensions. The results show that an appropriate selection of STFT parameters and the type of loss function in the DL model can improve the quality of the RIRs inferred.

1. INTRODUCTION

Artificial Intelligence (AI) has a wide application field in engineering, such as in robotics [1], image generation [2] and in computer vision. A field studied in AI is acoustics, with applications such as synthesis of natural speech from written language, voice modulation, and audio encryption. There has been recent interest in developing digital acoustic environments to provide a good immersion experience in virtual or augmented reality applications. However, the current available tools need high computing power and expensive investment in high quality equipment to perform as expected. In this paper we investigate the synthesis of a particular room impulse response (RIR) between two locations of a room by means of deep learning (DL) models and a set of measured RIRs from the same room.

Signal recorded by a microphone at a determined position inside a room ($s(n)$) can be modelled by the convolution (1) of the acoustic waveform originated by the source located at a specific position inside the room ($s(n)$) with the room impulse response (RIR) modelled as a linear time-invariant system ($h(n)$) of $L+1$ coefficients (2).

$$y(n) = s(n) * h(n) \tag{1}$$

$$h(n) = h_0\delta(n) + h_1\delta(n-\tau_1) + h_2\delta(n-\tau_2) + \cdots + + h_L\delta(n-\tau_L) \tag{2}$$

This model of Room Impulse Response [3,4] has been extensively used in the literature. In general, three parts can be distinguished in a RIR, as shown in Fig. 1. The first impulse is the direct path from the source to the microphone, later the first reflections arrive, those that bounce off the walls closer to the location of the microphone and usually denoted as “early reflection”, and finally, there is the late reverberation part. The measurement and quantification of the RIRs allow the creation of virtual acoustic spaces, in which using a clean sound source convolved with these RIRs would lead to virtually localizing this sound within the room [5].

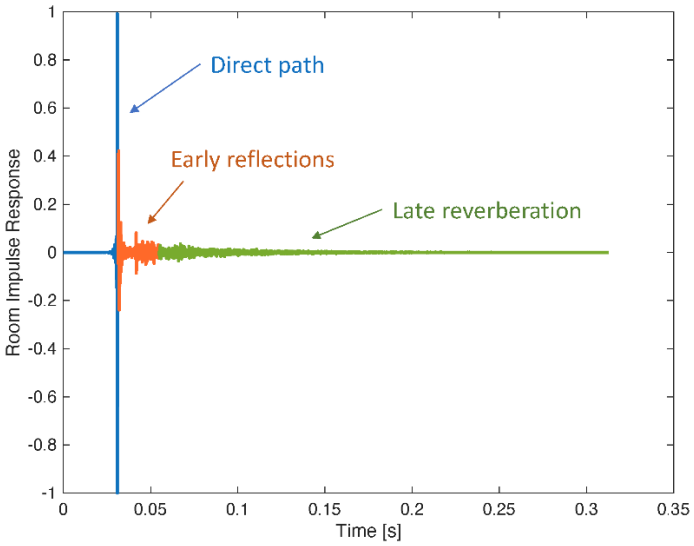

Fig 1. Different parts of a RIR. [3]

Modeling methods are usually classified into three categories: physical models, scale models and computational models. An often-used method is the IMS (Image Source Method) [6], where the concept of audio wave is replaced by that of audio ray. The method is based on the geometric construction of a specular reflection, copying the source in the plane that reflects the sound. Lately, some efficient methods to synthesize RIRs following IMS theory are available in repositories, as the FastRIR [7] and the gpuRIR tools [8].

Regarding the use of DL models to solve acoustic problems, one of the most used representations of the sound is the Short Time Fourier Transform, or STFT [9]. The STFT of an audio signal takes the waveform with its linear-time characteristics and converts them into a function of time and frequency. An example of a STFT can be seen in Fig. 2. This is calculated based on a time interval, called the hop size. Taking a frequency window size, and applying it to n inputs per window, we get the short-time Fourier transform (3).

$$X(m,k) = \sum_{n=0}^{N-1} w(n) x_m(n) e^{-j\frac{2\pi}{N}nk} \ , \qquad (3)$$

Where m is the frame index, n is the time index, k is the frequency index such that $0 \leq k < N-1$, where N is the size of the Fast Fourier transform (FFT). The window function $w(n)$ multiplies the m-th frame of signal $x(n)$ denoted by $x_m(n)$ in (3). The length of the frame is selected according to the required resolution: short frames give good resolution in time and long frames give good resolution in frequency. Usually, the frames are taken with a percentage of overlap between them, and the FFT size is at least twice the size of the frame. In this work we have used the STFT of the RIRs to feed the DL models. Figure 2 shows an example of the waveform of a RIR in time (top) and the magnitude of its corresponding STFT (bottom).

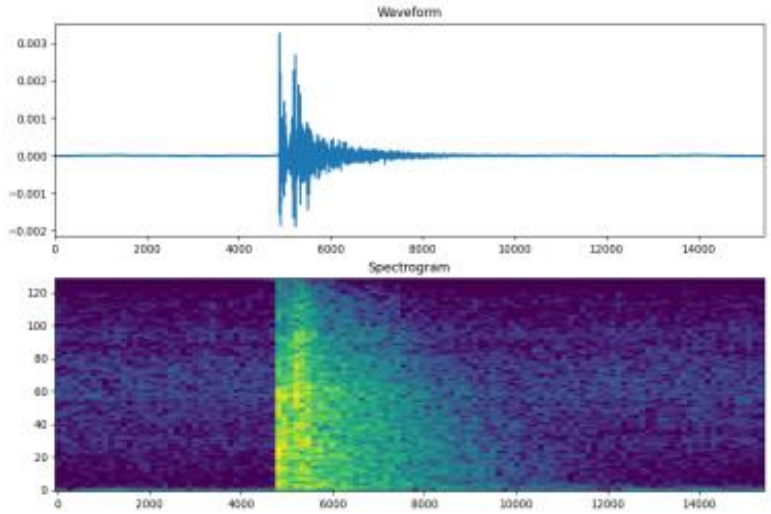

Fig 2. RIR waveform in time (top) and its corresponding STFT magnitude (bottom).

2. DEEP LEARNING APPROACH

Our DL model will make use of the information provided by the STFT of a singular RIR measured between the source and one specific location inside the room to obtain a new RIR between the source and another location. This new RIR does not belong to the training dataset, thus it is a new RIR inferred by the DL model. A further objective of this work is to transfer the DL model obtained for one specific room to infer RIRs of another room with different acoustic characteristics.

The DL model is based on the concept of Autoencoder (AE). This kind of model is defined by the compression and decompression of data for its regeneration, and it is mostly used to reconstruct images. Their applications are numerous, from image noise removal to out-of-range data detection, with applications to the safety control of nuclear power plants or air transport, where there are countless variables, and it is difficult to monitor all of them at the same time to detect a mistake.

The basic architecture of the AE is made up of three elements: an encoder that reduces the dimensionality of the data by increasing its depth, a bottleneck, where all the information is compressed into a vector of l terms, and a decoder, which increases and decrypts the latent vector to generate an image or data. The basic AE is made by means of convolutional layers [10].

For our problem, we have also used the Variational Autoencoder (VAE), which differs from the AE in its latent space and loss function. In the VAE, before logging the flattened data at the bottleneck, they are standardized into a normal Gaussian distribution. In this way, the VAE manages to create a more consistent and better distributed latent space. Figure 3 shows the classical structure of a VAE where the input and the output are images.

Finally, we have also used a third autoencoder that implements residual connections (Residual Autoencoder, ResAE), which provides a faster and more complete learning compared to the AE that uses only convolutional layers. The residual connections are created using the identity block, which does not perform any transformation and just maps the identity function, and the convolutional block, which maps the identity function and applies a convolution.

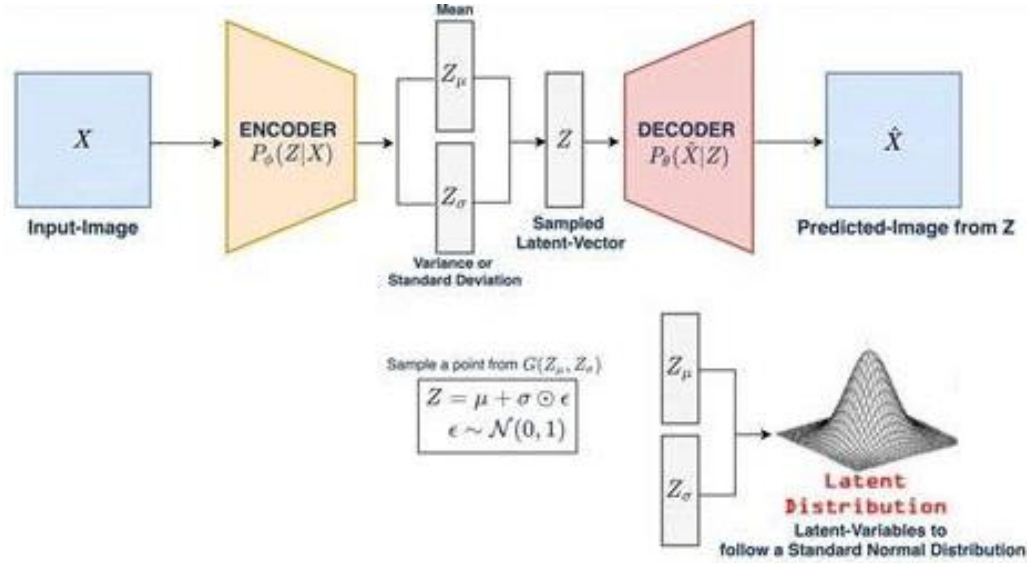

Fig 3. Structure of a Variational Autoencoder. [11]

2.1 Input to the Neural Network

One of the inputs to the neural network is the magnitude and phase of the STFT of a measured RIR. The STFT is meant to obtain accurate information in time, thus its window length is quite short (4ms). The specific parameters for a sampling frequency of $f_s = 16000$ Hz are:

- Size of the FFT (N): 256
- Window length: 64
- Hop length: 32 (overlap of 50%)

The STFT is a matrix of complex values of dimension $[N \times N_{frames}]$. Therefore, their magnitude and phase will be processed separately by the neural network. Moreover, decimal logarithm is applied to the STFT magnitude to obtain its power in decibels. These calculations are performed to attain more comprehensive data, since the simple transform results in reduced and not very extensive values in the data matrix. Normalization is performed using the minimum maximum scaling method, resulting in a normalization between 0 and 1. Figure 4 shows the magnitude (top) and the phase (bottom) of the STFT of a RIR.

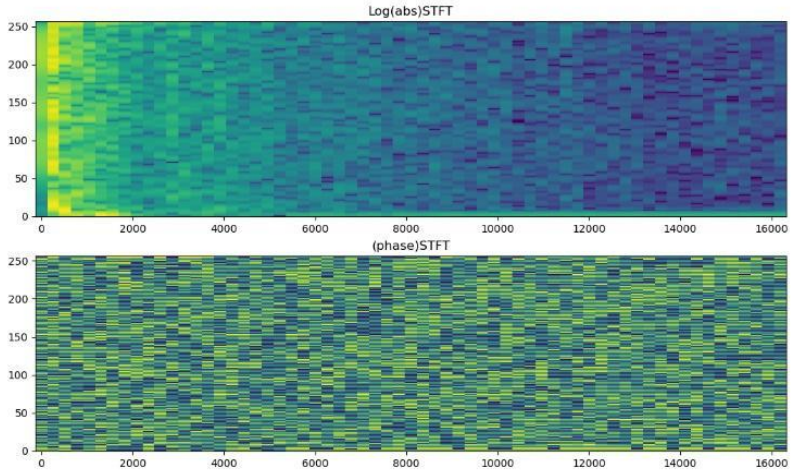

Fig. 4. Magnitude (top) and pase (bottom) of the STFT of a RIR.

The other input to the neural network is formed by two information vectors. We called information vector to a vector containing information about the locations of the source and the listener inside the room, together with the dimensions of the room and its reverberation time, similar to the input proposed in Fast-RIR [7]. Therefore, the information vector is formed by 10 parameters:

[lpx, lpy, lpz, spx, spy, spz, rdx, rdy, rdz, t60],

where lp{x,y,z} are the Cartesian coordinates of the listener and sp{x,y,z} are the Cartesian coordinates of the source, whereas rd{x,y,z} are the dimensions of the room and t60 is the reverberation time in seconds. All the length measurements are expressed in meters. Finally, the other input to the network consists of the information vector of the measured RIR (whose STFT feeds the network as well), stacked with the information vector of the RIR to be inferred.

2.2 Deep Learning Model

The model consists of 4 convolutional layers in the encoder and another 4 in the decoder, with the number of filters indicated in a list of convolution filters, kernels, and strides respectively, a latent space Z and l activation neurons for the latent vector. Each convolution layer is then applied a BatchNormalization layer and a ReLU activation, whereas in the decoder it is used a LeakyReLU activation. Figure 5 shows the structure of the VAE, where each color represents a residual layer.

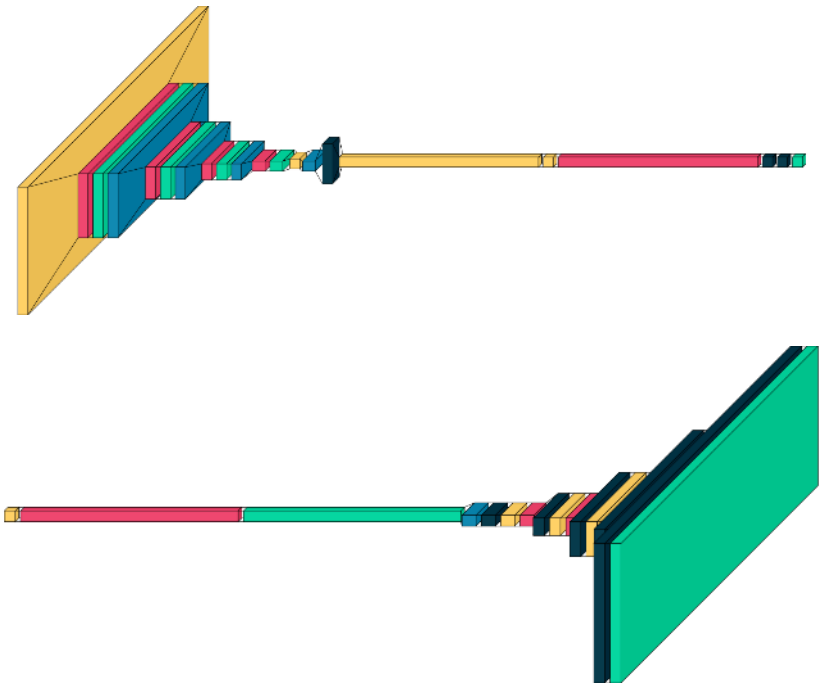

Fig. 5. Variational Autoencoder Architecture. Each color means a residual layer.

After applying a layer of l neurons to the latent vector and completely flattening the STFT data with the Flatten method, they are concatenated before being introduced into the latent space. Subsequently, the transposed convolutions with their BN are performed to obtain the output. The reconstruction error of the autoencoders is the MSE (4).

$$MSE = L(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N}(y - \hat{y}_i)^2 \quad (4)$$

However, in the case of VAE an additional term, the Kullback-Leibler divergence, or KL loss, is also taken into account in the loss function [12]. This information-theory equation quantifies the

proximity of two probability distributions. To obtain the loss function, the equation is simplified by comparing the probability distribution with a Gaussian one. The equation would then look like:

$$L_{KL} = -0.5 \sum_{i=1}^{N} \left(1 + \log\left(\sigma_{Z,i}\right) - \mu_{Z,i}^2 - \sigma_{Z,i}^2\right) \quad (5)$$

2.3 Database

The availability of databases containing enough RIRs measured to train a model is scarce [13]. Therefore, in this work we use the Fast-RIR tool [7] to simulate a rectangular room with different listening positions and sound sources, resulting in 360,000 RIRs. The parameters of the room are as follows:

- Room size [9 x 6 x 2.5] [m]
- T60 = 0.2 [s]
- Distance between positions 0.3 [m]

With these parameters, the impulse responses are generated, which are after preprocessed and associated with their information vector.

2.4 Training

The chosen learning rate is 10^{-5}, a higher learning rate resulted in models learning too fast and creating incorrect patterns of the data provided. For training, an Early Stopping criterium has been used, checking the losses in the validation set with a patience of 40 epochs. A batch size of 512 and a maximum of 1000 training epochs have been set. The loss function used is the MSE (4) algorithm with the addition of the KL loss (5), and the training optimizer used is the Adam gradient optimizer. For the training process, “InverseTimeDecay” is used to exponentially reduce the learning rate every 100 epochs.

Generation: As said before, the features that feed the model are the magnitude and phase of the STFT and the two information vectors. However, the output of the model should be only the magnitude and phase of the STFT of the inferred (new) RIR. Once the new STFT has been generated by the DL model, it is necessary to de-normalize it. It is also necessary to undo the operations of the STFT and calculate its inverse STFT to obtain the waveform in time. Lastly, the audio wave is saved in .WAV format and its STFT is saved in two .NPY files, one for its magnitude in dB and the other for its phase.

3. RESULTS

3.1. Training Loss Results

We have compared three autoencoders: the basic autoencoder (AE), the variational autoencoder (VAE), and the residual autoencoder (ResAE). Table I shows the loss for each model for the training and validation data, as well as the number of epochs, the training time and the number of parameters of each model. It can be concluded from the training data that neither a long training time nor an exceeding number of parameters improved the models’ efficiency. Furthermore, the models converged similarly, except for the Autoencoder which trained for too long due to technical difficulties. Comparing results and parameters of the models it can be observed that the Variational Autoencoder outperforms the other two models, and it will be selected to be improved in future advances. Although a new training with a reduced set of parameters for the Residual Autoencoder will be performed before choosing the Variational Autoencoder.

It can be noted from Table I that the results from the Autoencoder training obtained a lower validation loss than training loss. This is not an expected result, meaning that the model performed better in inputs that have not trained on than the ones that has experienced. For the Variational

Autoencoder, it obtained the best results among all models. The Gaussian distribution created a uniform latent space and, therefore, it creates more consistent results. It also has the smaller number of parameters and a shorter training time. The results obtained by the Residual Autoencoder may point out that the selection of the complexity for the model (number of parameters) was excessive, and that a lighter model would have performed better.

TABLE 1. LOSS OBTAINED AT TRAINING.

Model	**Loss**	**Val-Loss**	**Epochs**	**Time [h]**	**N-params**
AE	$15.15 \cdot 10^{-4}$	$1.341 \cdot 10^{-3}$	1000	79	$3.5 \cdot 10^{5}$
VAE	$12.93 \cdot 10^{-4}$	$1.443 \cdot 10^{-3}$	111	23	$3.2 \cdot 10^{5}$
ResAE	$16.25 \cdot 10^{-4}$	$2.088 \cdot 10^{-3}$	111	19	$6.1 \cdot 10^{5}$

3.2 Generation Loss Results

Figure 6 shows the waveform of the real (top) and generated (bottom) waveform of the desired RIR. It can be seen how the Variational Autoencoder obtains a good estimation of the new RIR regarding the arrival of the direct sound at the right time. However, the early reflections are over-factorized. The STFTs of the same RIRs are shown in Figure 7, where the magnitude and phase of the STFT of the real and generated RIRs are displayed at the top and the bottom of the figure respectively. It can also be noted the effect of the over estimation of the early reflections in the magnitude plots. Regarding the phase plots, it can be noted some disturbance in the first miliseconds of the RIR and along the late reverberation.

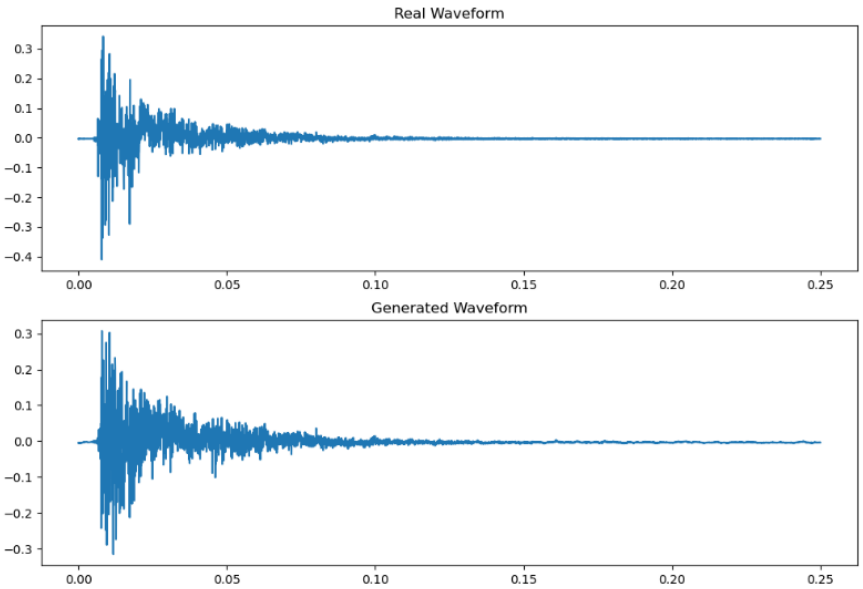

Fig. 6. Comparison between generated and simulated RIR.

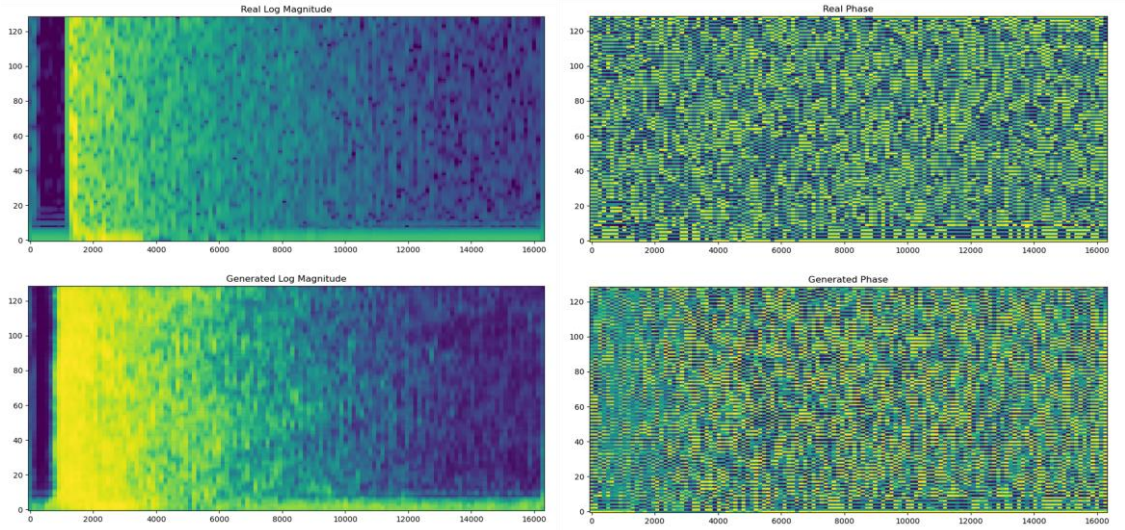

Fig. 7. Comparison between generated and simulated STFTs.

Table 2 shows the mean square error (MSE) of the three models for the whole RIR, as well as considering only their first 50ms, whereas Table 3 shows the MSE of the STFT for both magnitude and phase. The test set is formed by 21 generated RIRs. It can be noted from both tables that the MSE is very similar for the AE and the VAE, being much lower than that of the ResAE. Regarding the MSE of the RIR for the first 50ms, it is higher than for the full RIR, which is comprehensible result since most of their energy is concentrated within the first 50ms. However, the performance of the AE and VAE are always better than that of the ResAE, presenting an MSE seven times higher compared to the values obtained for the AE and the VAE, and for both domains, time and frequency.

TABLE 2. RESULTS OF RIR MEAN SQUARE ERROR

Model	MSE	MSE (50ms)
AE	1.073 · 10-3	5.285 · 10-3
VAE	1.084 · 10-3	5.329· 10-3
ResAE	7.867 · 10-3	35.984 · 10-3

TABLE 3. RESULTS OF STFT MEAN SQUARE ERROR

Model	**MSE Log Mag**	**MSE Phase**
AE	4.030 · 10-3	33.425 · 10-3
VAE	4.516 · 10-3	34.349· 10-3
ResAE	26.639 · 10-3	71.508 · 10-3

Results in Table 3 indicate that most of the MSE achieved by the models comes primarily from the reconstruction of the phase. The STFT magnitude can be predicted reasonably well, while the STFT phase obtains errors eight times higher than the magnitude.

3.3 Model Execution Times

Next, it will be commented the execution times for the generation and postprocessing of the different models. Table 4 shows the generation time and the postprocessing time in seconds for all the models for the whole test set of 21 RIRs.

TABLE 4. EXECUTION TIME FOR A BATCH OF 21 RIR

Model	Generation time (s)	Post Process time (s)
AE	1.407	0.307
VAE	2.149	0.049
ResAE	2.671	0.050

The Autoencoder model took the less time to generate the RIR, taking $6.7 \cdot 10^{-2}$ seconds to generate each RIR. The Variational Autoencoder took approximately 0.6 seconds more than the AE for the generation time since the latent space is encoded in a Gaussian distribution and needs some additional computation. For the Residual Autoencoder, as the number of parameters is doubled, it takes 1.2 seconds more than the AE. The post processing times indicate that the first postprocessing takes slightly more time, but it could be overlapped with the generation process.

4. CONCLUSIONS

The problem of inferring a new RIR from a measured RIR in a seam room has been studied. Three neural networks based on the Autoencoder structure have been implemented to generate the new RIR, whose input was the same and consisted in the STFT of the measured RIR together with two information vectors, one referred to the measured RIR position, and the other referred to the new RIR's. Apart from this information, the vectors included the dimensions of the room and its reverberation time, in case the resulting model would be used for a different room. The results obtained satisfied the need for the problem to solve, although the underperformance of the Residual Autoencoder leaves the incognita if a better choice of hyperparameters and weight model would lead to better results. It can be observed how the Variational Autoencoder presents the best performance regarding the losses, but the MSE is similar to that of the Autoencoder for the test set. Nevertheless, the VAE obtains realistic Room Impulse Responses, especially considering the arrival time of the direct path.

5. FUTURE WORK

The future work, which is aimed to improve the system, includes the implementation of a loss function adapted to the STFT as in [14], or, alternatively, to separately train two models, one for the magnitude and one for the phase. Other possibility is to trim the RIR such that the direct path and early reflections are trained with a model and the late reverberation with a different one, due to their different statistical properties. Last but not least, other DL models could be used such as adversary networks or transformers.

ACKNOWLEDGEMENTS

This work was supported by the Spanish Ministry of Science and Innovation through Grant No. PID2021-124280OB-C21 (MCIU/AEI/FEDER, UE).

REFERENCES

[1] M. Brady. Artificial intelligence and robotics, in Robotics and Artificial Intelligence, p. 47-63, 1984.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint, vol. arXiv:2204.06125, 2022.

[3] P. A. Naylor and N. D. Gaubitch, Speech dereverberation, London: Springer, 2010.

[4] C. Störig and C. Pörschmann, Investigations into velocity and distance perception based on different types of moving sound sources with respect to auditory virtual environments, vol. 10, JVRB-Journal of Virtual Reality and Broadcasting, 2014.

[5] Z. Chen and R. Maher, "Addressing the discrepancy between measured and modeled impulse responses for small rooms," in Audio Engineering Society Convention 123, 2007.

[6] P. Mehta and V. Bhadradiya, "Measurement of room impulse response using image source method," in 2015 International Conference on Electrical, Electronics and Mechatronics, 2015.

[7] Ratnarajah, A., Zhang, S.-X., Yu, M., Tang, Z., Manocha, D., & Yu, D., "Fast-RIR: Fast Neural Diffuse Room Impulse Response Generator", in *Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 571–575, 2022

[8] Diaz-Guerra, D., Miguel, A., & Beltran, J. R., "gpuRIR: A Python Library for Room Impulse Response Simulation with GPU Acceleration", *Multimedia Tools and Applications*, *80*(4), pp.5653–5671, 2018.

[9] J. A. Moorer, " A note on the implementation of audio processing by short-term Fourier transform," in 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017.

[10] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biol. Cybernetics, vol. 36, p. 193–202, 1980.

[11] "Matworks," 2022. [Online]. Available: https://la.mathworks.com/discovery/autoencoder.html. [Accessed 18 07 2022].

[12] J. Shlens, "Notes on kullback-leibler divergence and likelihood," arXiv preprint, vol. arXiv:1404.2000, 2014.

[13] D. D. Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge and S. Gannot, "dEchorate: a calibrated room impulse response dataset for echo-aware signal processing," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2021, no. 1, pp. 1-15, 2021.

[14] 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), 2020.