DOI: 10.61782/fa.2023.1177

forum acusticum 2023

GENERATING IMPULSE RESPONSES USING AUTOENCODERS

Martin Eineborg
Treble Technologies
me@treble.tech

Finnur Pind
Treble Technologies
fp@treble.tech

Solvi Thrastarson
Treble Technologies
sth@treble.tech

ABSTRACT

Room Impulse Response (IR) captures the characteristics of the associated room and is affected by geometry, boundary materials, and objects within the room. Being able to generate IR signals for any given position in a room is a complicated problem. Storing all data points to cover a space, in order to be able to seamlessly move around, carries large data storage requirements. In this work we examine using autoencoders and neural networks as an efficient way to store data points and interpolate to new locations. The results is a generation of IRs in near real time that can be used in many scenarios e.g. moving around in games and virtual worlds, as well as for giving users the possibility of walking around in a building and experience an auralization of the soundscape.

A set of IR signals was created by an acoustic simulation engine using a source and multiple receivers. The data set was then used to train an autoencoder to compress IR signals to a latent space representation which was used to train a fully connected multi-layer neural network to generate IR signals for any given position in the room.

The result shows that high fidelity IR signals can be predicted with significant reduction in storage size using autoencoders and neural networks.

Keywords: *Impulse Responses (IR), Autoencoder, Neural Networks*

1. INTRODUCTION

Room Impulse Response (IR) can be measured at any given point in a room for a source location, and it captures the acoustic characteristics of the room. It is affected by the geometry of the room, as well as materials on surfaces and objects in the room. IR signals can be used to optimize microphones, speakers, acoustics of buildings, concert halls, etc. Being able to compute IR signals for any given position in a room is a computationally demanding task and storing IR signals for all positions in a room is very memory consuming so interpolating between locations with stored IRs to estimate the IRs in between is a common approach.

There has been much research on using machine learning to create IR signals. For example, [1] makes the case for using a neural network to interpolate IR signals and as an efficient way to store IR signals. Generating IRs directly from images of rooms have been explored by e.g. [2] who uses Generative Adversarial Networks (GAN) and [3] makes use of Transformers. [4] uses a GAN to generate IRs in unseen environments and with changing parameters e.g. reverberation time. Generating IR using a Variational Autoencoder was explored in [5].

The work in this paper is focused on using Autoencoders, introduced by [6], to compress IRs and then using a neural network to predict, for any given location, the corresponding IR. The problem is to learn $e : R^n \to R^p$ and $d : R^p \to R^n$ such that

$$\arg\min_{d,e} E[d(x, d(e(x)))].$$

The informal idea behind an autoencoder network is to train a network to reproduce the input by reconstructing it from an internal representation of much lower dimension. This internal latent space representation can then be utilized as a way to represent the original signal in a compressed way.

2. THE EXPERIMENTS

In the following chapters we discuss the proposed method, the data used for the experiments, as well as the machine learning methods that were used.

2.1 The Proposed Method

The experiments have been conducted in two phases. During the first phase, an autoencoder was trained on IRs in order to compress the IRs and then use the encoder and decoder to transform to and back from latent space representation, respectively. In the second phase a neural network was trained on room positions to predict corresponding latent space representations (of the autoencoder).

2.2 Synthetic Data Set

For the experiment we used synthetic IR signals created from an audio simulation tool, the Treble Acoustic Simulation Suite, for an L-shaped room (see Fig 1) consisting of 1 source and 195 receivers in a 2-D plane. The simulations were done using the Discontinuous Galerkin Method (DGM) [7] which has been developed to solve wave-equation-based problems [8]. The solver we used is GPU-accelerated and developed to solve the acoustic wave-equation for various room acoustics scenarios [9] [10].

We compute an impulse length of $1.2s$, with an upper frequency of 1420 Hz, for an L-shaped room made up of two $3x7m^2$ boxes (see Fig. 1). In this work, only the first $0.1s$ of the IR was used. The data was split into training and validation sets using an 80/20 split. The first 2 000 data points of each IR signal was used for this initial feasibility study but we aim to extend it to full IRs in future work.

The experiment was reproduced with the source moved to another location in the same room with the same receivers. The result is built on both data sets. The source positions are $(1.5, 0.5, 1.5)$ and $(1.5, 5.5, 1.5)$ with origin being upper right corner.

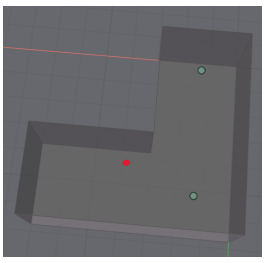

Figure 1. The room used for generation of synthetic IR signals. The green points shows locations of the two sources used in the experiments. The red point is the receiver from Fig 7.

2.3 Autoencoder

An autoencoder was trained on the examples, each consisting of 2 000 data points. The encoder part, transforming the example dimension of 2 000 to a latent space of 128, can be seen in Fig 2 (left) below.

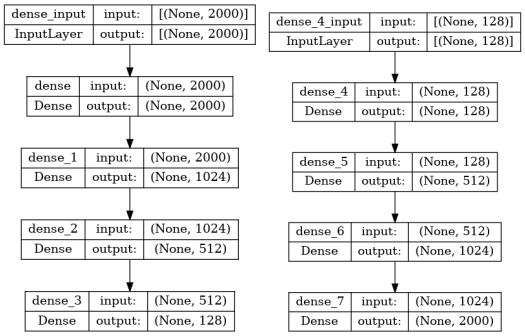

Figure 2. The autoencoder architecture with an encoder (left) and a decoder (right)

The decoder part of the autoencoder, responsible for transforming latent space representation back to an IR signal, can be seen in Fig 2 (right). The layers of the autoencoder use rectified linear activation function, the Nadam optimizer [11], a learning rate of 0.001, and a batch size of 64.

2.4 Neural Network for Predicting IRs

An artificial neural network of 7 layers was trained on X, Y coordinates in the room (Z is fixed) to predict a latent space representation. It consisted of layers varying between 2 048 and 256 neurons. It was trained for 5 000 epochs on the same data as the autoencoder but now transformed into latent space. The predicted representation was then transformed to the corresponding IR signal using the decoder part of the autoencoder.

3. RESULTS

The autoencoder was trained using 10-fold cross-validation for 1 000 epochs and the performance on each validation set can be seen in Fig 4 below. The training resulted in a Mean Squared Error (MSE) of the validation set of $1.4e-05$. The dimension of IR signals has been reduced by 93.6% by using the latent space as a representation.

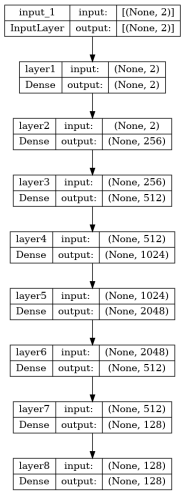

Figure 3. Neural Network Architecture

Both the autoencoder and the fully-connected feed-forward neural network was constructed using Tensorflow [12] and Keras [13].

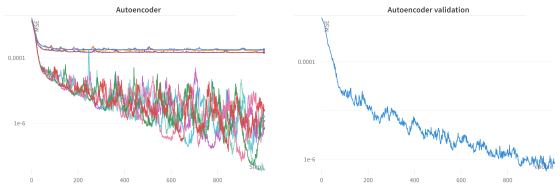

Figure 4. MSE of the 10 different validation datasets in the 10-fold cross-validation during training of the autoencoder (left) and the median value of the 10 folds (right).

An example of how an unseen IR signal which has been passed through the autoencoder looks like compared to the original IR signal can be seen in Fig 5. The graph shows the matching of the target and the reconstructed IR signals. The red dashed line shows the reconstructed IR signal and the blue line shows the target high fidelity IR signal.

The fully connected forward feed neural network responsible for predicting latent space representations of IR signals using room coordinates was trained for 5 000 epochs with a resulting MSE of $1.09e-4$ for the validation set. The training and the performance on the training and evaluation set can be seen in Fig 6 where MSE is displayed on a log scale.

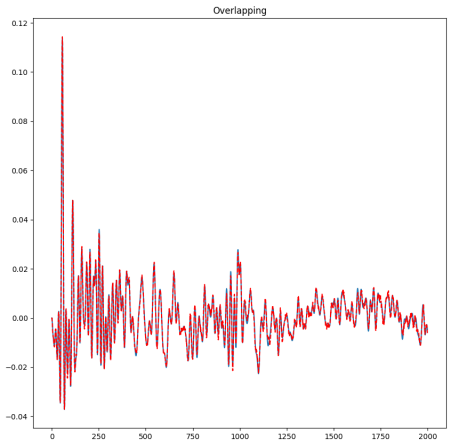

Figure 5. Matching of IRs. The red dashed line is the reconstructed IR, the blue is the target IR

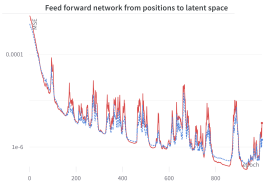

Figure 6. MSE of the training (red line) and validation (blue dotted line) of the feed-forward neural network

The IR signal of an unseen receiver was generated by feeding it's coordinates to the neural network. The location of the receiver can be seen in Fig 1. The decoder part of the autoencoder was then used to return the corresponding IR signal, Fig 7. The graph shows both the target and the predicted IR signals. The blue line is the predicted IR signal and the red dashed line is the target IR signal.

	Autoencoder	Neural Network
MSE	$1.4e-05$	$1.09e-04$

Table 1. MSE

4. CONCLUSION

The proposed method of using autoencoders to compress IR signals and use neural network to interpolate IR signals

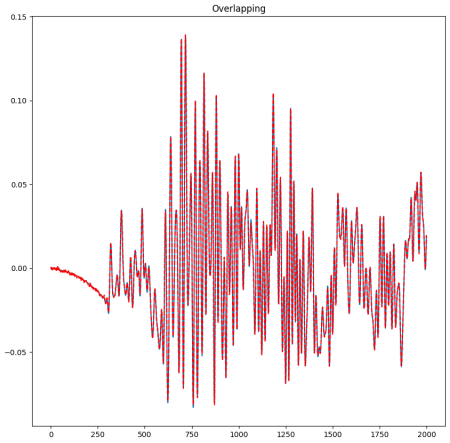

Figure 7. Matching of IRs. The red dashed line is the generated IR, the blue is the target IR.

in rooms trained using synthetic data was shown to be an promising way to interpolate and generate new IR signals both for already seen and previously unseen locations in the rooms.

5. ACKNOWLEDGEMENTS

The authors would like to thank the entire team at Treble Technologies.

6. REFERENCES

[1] A. Richard, P. S. Dodds, and V. K. Ithapu, "Deep impulse responses: Estimating and parameterizing filters with deep networks," *CoRR*, vol. abs/2202.03416, 2022.

[2] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori, "Image2reverb: Cross-modal reverb impulse response synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 286–295, October 2021.

[3] S. Majumder, C. Chen, Z. Al-Halah, and K. Grauman, "Few-shot audio-visual learning of environment acoustics," 2022.

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

[5] D. Sanaguano-Moreno, J. Lucio-Naranjo, R. Tenenbaum, L. Bravo-Moncayo, and G. Regattiere-Sampaio, "A deep learning approach for the generation of room impulse responses," in *2022 Third International Conference on Information Systems and Software Technologies (ICI2ST)*.

[6] G. H. D.E. Rumelhart and R. Williams, *Learning internal representations by error propagation. In Parallel Distributed Processing*. Cambridge, MA: MIT Press, 1986.

[7] W. H. Reed and T. R. Hill, "Triangular mesh methods for the neutron transport equation," tech. rep., Los Alamos Scientific Lab., N. Mex.(USA), 1973.

[8] M. Käser and M. Dumbser, "An arbitrary high-order discontinuous galerkin method for elastic waves on unstructured meshes—i. the two-dimensional isotropic case with external source terms," *Geophysical Journal International*, vol. 166, no. 2, pp. 855–877, 2006.

[9] F. Pind, C.-H. Jeong, A. P. Engsig-Karup, J. S. Hesthaven, and J. Strømann-Andersen, "Time-domain room acoustic simulations with extended-reacting porous absorbers using the discontinuous galerkin method," *The Journal of the Acoustical Society of America*, vol. 148, no. 5, pp. 2851–2863, 2020.

[10] A. Melander, E. Strøm, F. Pind, A. Engsig-Karup, C. Jeong, T. Warburton, N. Chalmers, and J. Hesthaven, "Massive parallel nodal discontinuous galerkin finite element method simulator for room acoustics," in *International Journal of High Performance Computing Applications*.

[11] T. Dozat, "Incorporating Nesterov Momentum into Adam," in *Proceedings of the 4th International Conference on Learning Representations*, pp. 1–4.

[12] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," 2016.

[13] N. Ketkar and N. Ketkar, "Introduction to keras," *Deep learning with python: a hands-on introduction*, pp. 97–111, 2017.