DOI: 10.61782/fa.2023.0646

REAL-TIME AURALIZATION BASED ON A CONVOLUTIONAL NEURAL NETWORK TRAINED BY LOW-FREQUENCY WAVE-BASED CALCULATIONS

Dingding Xie[1,2*] **Maarten Hornikx**[1]

[1] Department of the Built Environment, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands;

[2] National Key Laboratory of Underwater Acoustic Technology, Harbin Engineering University, 150001 Harbin, China.

ABSTRACT

Acoustic Virtual Reality (AVR) has been increasingly used in building design, where sound fields are expected to be updated in real-time as the receiver and source move around the space. At low frequencies, wave-based methods can be used in the pre-calculation stage to obtain credible sound fields, that are stored for later real-time interpolation, which may lead to large storage requirements. This research proposed Neural Networks (NNs) trained by results from low-frequency room acoustic calculations such that they can provide the binaural room impulse responses (BRIRs) in real-time in an AVR framework. The room sound fields were calculated by solving the Helmholtz equation through the Finite Element Method and stored at spherical receiver arrays to build the training datasets. Convolutional Neural Networks are used to predict the spherical harmonics (SH) coefficients of the sound field distribution on spherical receiver arrays with the positions of the source and receiver as input. Combined with head-related transfer functions, these SH coefficients can be used to obtain BRIRs efficiently. At the cost of training NNs, this method is applicable to AVR scenarios with moving source and receiver and arbitrary head orientation, with the advantages of fast real-time calculation and distinct storage data reduction.

Keywords: *acoustic virtual reality, convolutional neural networks, room acoustics, low frequencies.*

1. INTRODUCTION

In recent years, AVR has been increasingly used in architecture and building design to obtain an acoustic perception that is close to reality. Since the room impulse responses (IRs) are dependent on the source and receiver position, sound fields should be continuously updated in real-time as the receiver and source move around in the space, and the spatialization process must be recalculated as the head orientation changes.[1] The calculation approaches can be divided into two main categories: the real-time calculation approach and the pre-calculated approach. The real-time calculation approach applies simplifications to get real-time performance within the strict time, which will naturally decrease the accuracy. The pre-calculated approach finishes most computations in a pre-calculation stage. At low frequencies, as all relevant wave behavior has an important influence on sound field composition,[2] wave-based methods are generally necessary to obtain credible sound fields. Then the simulated IRs should be stored on the grid with spatial information.[1] During run-time, interpolation is done to obtain the sound field distributions on spherical receiver arrays corresponding to certain transducer positions. Plane-wave density functions (PWDs) can be calculated using plane wave decomposition from the sound field distributions. By convolving the PWDs with the head-related transfer functions (HRTFs), BRIRs including the impacts of head

[*] *Corresponding author*: d.xie@tue.nl.

and room can be obtained for auralization.[3] This approach leads to the problem of large storage requirements,[1] and the challenge of computing the plane-wave decomposition at interactive rates.[4]

Machine Learning (ML) in acoustics is rapidly developing in recent years.[5] In the field of room acoustics, Pulkki *et al.* proposed fitting NNs with the input of geometric features and output of the filter parameters to render the acoustic effect of scattering from finite objects and provide a perceptually plausible response for the listener.[6] Tenenbaum *et al.* proposed a methodology using a radial basis functions type of artificial NNs trained by the BRIRs patterns to save computational time spent on the classical convolution method and produce faster auralization.[7] Fernandez-Grande *et al.* proposed generative adversarial networks to reconstruct sound fields from experimental data and recover some of the sound field energy that would otherwise be lost at high frequencies.[8] Notably, Borrel-Jensen *et al.* presented a physics-informed neural network to predict the solution to the linear wave equation to obtain the sound field in 1D with parameterized sources and impedance boundaries, and this method will be further applied in realistic 3D scenes.[9] In addition, there is a mesh-based neural network to generate impulse responses (IRs) for indoor 3D scenes whereby 3D scene meshes were transformed into latent space and the latent space was used to generate IRs,[10] and a Neural Acoustic Fields methodology that represents how sounds propagate in a physical scene and learns to continuously map all emitter and listener location pairs to a neural impulse response function.[11] Besides, NNs are used to predict reverberation time, room volume, absorption coefficients, and eigenfrequencies, and analyze multi-exponential sound energy decay.[12-16]

This paper presents the basis for real-time auralization: a method based on a Convolutional Neural Network (CNN) trained by low-frequency wave-based calculations, which can be used in real-time convolutions and is applicable to the scenario of sound fields with moving sources and receivers, and varying head orientation.

2. RESEARCH METHODS

For AVR purposes, the low-frequency sound fields as sampled and stored at spherical receiver arrays should be decomposed into plane waves using the spherical harmonics expansion. To reduce the storage requirements of sound field data and save the computation load of plane-wave decomposition, we propose an NNs-based method, with the positions of source and receiver as input and SH coefficients that vary with frequency as output. In this work, the dataset will be established by solving the Helmholtz equation by the Finite Element Method (FEM) which ensures the accuracy of the sound field simulation for elements sizes that have been chosen properly small. After the Networks are trained, we can obtain SH coefficients at any transducer positions within the training areas in a certain room. In future work, by combining with the SH coefficients of room sound field distributions and the SH coefficients of HRTFs, the BRIRs can be obtained efficiently in real-time,[3,17] as shown in Fig. 4.

2.1 Input of Neural Networks

The input datasets are composed of images showing source and receiver position distances to the room boundaries with pixels of 28×28. The distances from the sound source to walls in different directions are calculated and plotted in a sphere whose center is the source, and an example is shown in Fig.1(a) and mapped onto a plane as shown in Fig.1(b). Every value on the image shows the distance from the source to the wall in that direction. Images showing receiver positions are given in the same way, as shown in Fig.2. These images include information on transducer positions and room geometrics.

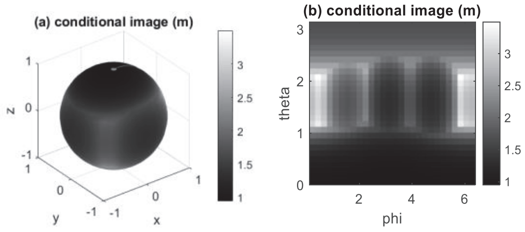

Figure 1. Images presenting source position relative to room boundaries.

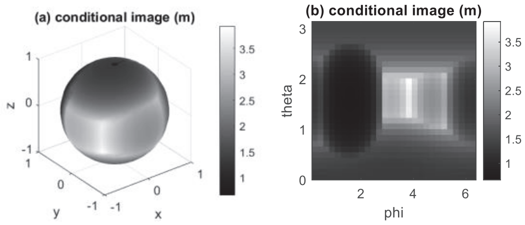

Figure 2. Images presenting the receiver position relative to room boundaries.

2.2 Output of Neural Networks

At low frequencies, a wave-based method was used to simulate the sound field in rooms. The sound field was excited by a point source, and a spherical microphone array was used to sample the sound pressure on a sphere. The equal-angle sampling method was used which means the azimuth angle ϕ and elevation angle θ are both sampled with the same number as shown in the following equations:

$$\theta_q = (q+\frac{1}{2})\frac{\pi}{2v+2}, q = 0,...,2v+1, \tag{1}$$

$$\phi_l = l\frac{2\pi}{2v+2}, l = 0,...,2v+1, \tag{2}$$

therefore, the total number of samples is given by $(2v+2)^2$.[18] Notably, the sampling distance should follow the spatial sampling theorem.

The sound field can be decomposed into plane waves using a spherical harmonics expansion. The SH coefficients A were calculated using:

$$A_{nm} = \frac{1}{j_n(ka)}\iint p(a,\theta,\phi)Y_n^m(\theta,\phi)^* \sin\theta d\theta d\phi, \tag{3}$$

where $p(a,\theta,\phi)$ is the sound pressure distribution on the sampling sphere with the radius of a, $k = 2\pi f / c$ is the wave number, $j_n(ka)$ is the spherical Bessel function of the first kind, and $Y_n^m(\theta,\phi)$ is the spherical harmonics function.[19]

With all orders of SH coefficient, the plane-wave density function $a(\theta,\phi)$ for every frequency can be obtained using:

$$a(\theta,\phi,f) = \sum_{n=0}^{N}\sum_{m=-n}^{n}\frac{A_{nm}(f)}{4\pi i^n}Y_n^m(\theta,\phi), \text{[3,19]} \tag{4}$$

where N is the highest order of spherical harmonics and is decided by:

$$ka < N . \text{[18]} \tag{5}$$

The output datasets were composed of the SH coefficients as a function of frequency for different orders of spherical harmonics, in real part and imaginary parts separately, and an example is given in Fig.3. These SH coefficients can be directly combined with HRTFs in future work which contributes to fast real-time calculation.

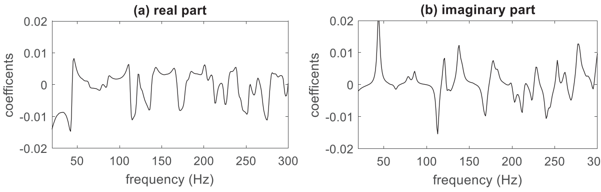

Figure 3. SH coefficient for monopole (a) real part (b) imaginary part.

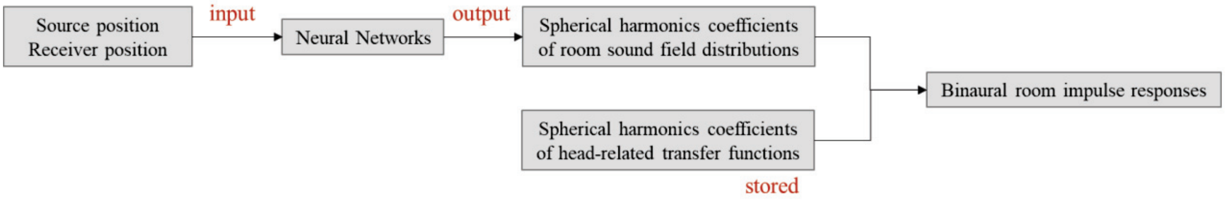
Source position
Receiver position
input
Neural Networks
output
Spherical harmonics coefficients of room sound field distributions
Binaural room impulse responses
Spherical harmonics coefficients of head-related transfer functions
stored

Figure 4. Research approach rationale.

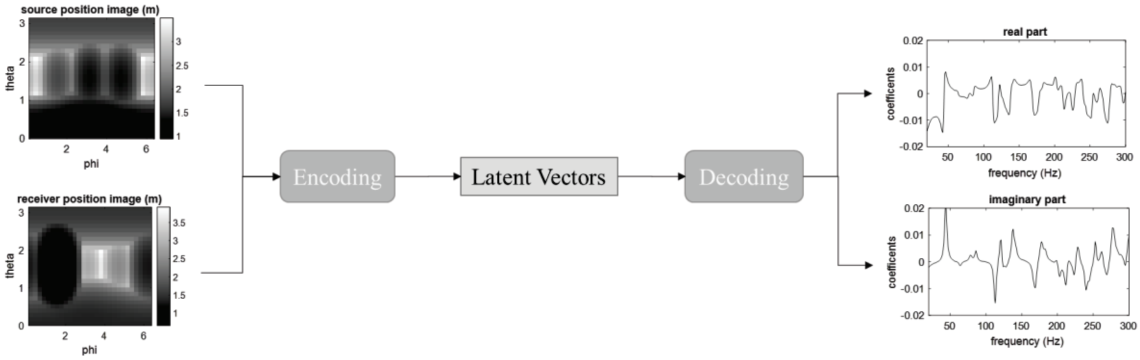

Figure 5. The structure of Neural Networks.

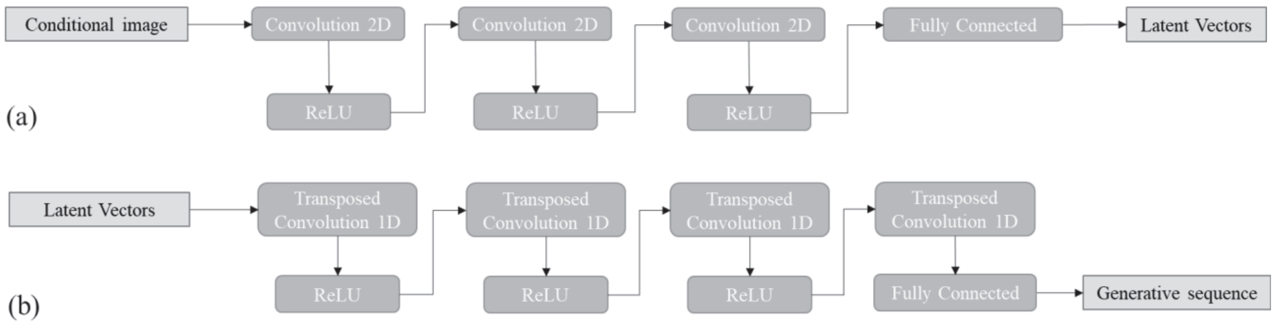

Figure 6. The architecture of the (a) Encoding Network (b) Decoding Network.

2.3 Structure of Neural Networks

Next, we used a CNN for image-to-sequence regression, a variant of Variational Autoencoder,[20] to learn the frequency-dependent SH coefficient curves. The structure of the NNs is shown in Fig.5. It is composed of two parts: Encoding Networks and Decoding Networks. The encoder takes an image input and outputs a Latent Vector representation using a series of down-sampling operations such as 2D convolutions as shown in Fig.6 (a). The input images of the NNs are the images presenting transducer positions. Latent Vectors are the suitable internal representation of input images. The decoder takes as input a Latent Vector and reconstructs the sequence using a series of up-sampling operations such as 1D transposed convolutions, as shown in Fig.6 (b). The output sequences are the curves presenting the SH coefficients varying with frequency. In Figures 6, the ReLU layer performs a threshold operation on each element of the input, where any value less than zero is set to zero.

3. BENCHMARK CALCULATIONS

3.1 Dataset building

To build the dataset, the COMSOL6.0 pressure-acoustics module was used to simulate the sound fields in rooms at low frequencies. The room model and sampling sphere with a radius of 20 cm are shown in Fig.7(a). There were 10 samples positioned along with both azimuth and elevation angle, 100 sampling positions in total. All boundaries are impedance boundaries.

In this example, the highest frequency is 300 Hz. Sound fields with 16×11 sound source positions (Fig.7(b)) were simulated and 16×11 receiver positions (the center of the receiver as shown in Fig.7(c)) were sampled. The distance between two nearest sound sources or receivers is one-fifth of the wavelength of the highest frequency, here it is 22.86 cm for 300 Hz. The size of the dataset (number of source and receiver combinations) is 30976.

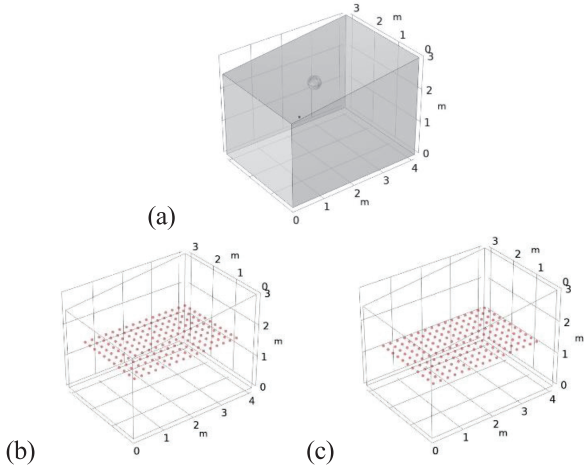

Figure 7. (a) Room model (b) Source positions (c) Receiver positions.

3.2 Neural Networks training

MATLAB was used to train the NNs. The input of the NNs is the transducer images in 2 channels. As the spherical harmonics are truncated at order 2 at 300 Hz, there are 9 coefficients in total and every coefficient is divided into two real and imaginary parts, there are 18 channels as output for this case.

The loss L measuring how close the decoder output is to the ground truth by using the root-mean-square deviation is given by,

$$L = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(x_{gj}-x_{rj})^2}\ , \qquad (6)$$

where m is the number of points, x_{gj} and x_{rj} are the values for every point for the generative sequences and real sequences respectively. Training options are shown in Table 1, where the mini-batch is a subset of the training set that is used to evaluate the gradient of the loss function and update the weights, epoch presents the number of times all of the training vectors in the mini-batch are used once to update the weights. Validation frequency gives how many subsets the training dataset contains and presents the iteration per epoch, therefore the total iteration is the product of validation frequency and epoch. Fig.8 gives the curve of loss varying with epoch. With increasing training, the loss is getting smaller which means the sequences generated are getting close to the real sequences.

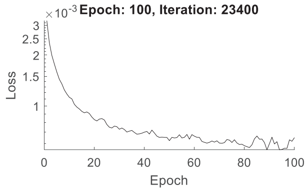

Figure 8. Loss curve (in the Y-log scale).

Table 1. Training options settings.

Training Options	
Training dataset size	30000
Test dataset size	976
Mini Batch Size	128
Validation frequency	234
Epoch	100
Iteration	23400

3.3 Results

To verify the performance of the NNs, randomly choosing one pair of source and receiver positions, Fig.9 shows the comparison of real sequences and generative sequences for the monopole. Accordingly, the sequences generated by the NNs are close to the real curves for both the real and imaginary parts.

For the frequency range from 20 to 300 Hz, Eq. (4) was used to calculate the PWD. For the frequency range of 20-224 Hz and 224-300 Hz, the spherical harmonic expansion should be truncated at order 1 and 2 respectively according to Eq. (5). At the frequency of 300 Hz, the real and predicted plane-wave density functions are shown in Fig.10(a) and (b).

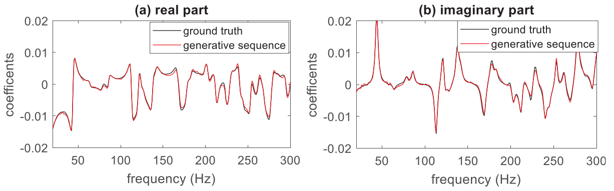

Figure 9. SH coefficients for the monopole (a) real part (b) imaginary part.

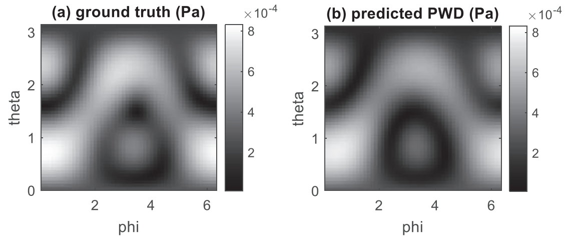

Figure 10. The absolute value of PWD at 300 Hz (a) ground truth (b) predicted value.

For the position of $\theta = \pi / 100$ and $\phi = \pi / 25$, filtered by a filter in Fig.11, the frequency spectrums are shown in Fig.12 in real and imaginary parts. As we can see, the results predicted by the NNs are close to the results from FEM which we use as ground truth.

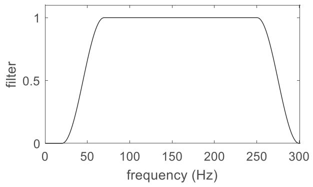

Figure 11. Magnitude response of the filter.

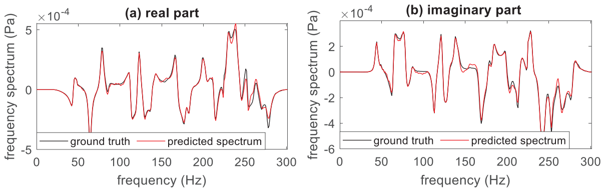

Figure 12. Comparison of real and predicted frequency spectrums (a) real part (b) imaginary part.

The inverse Fourier transform of the frequency spectrum is the IR, as shown in Fig.13(a). The error (difference) between the predicted response and ground truth is shown in Fig.13(b). The error is rather small compared with the ground truth which ensures the reliability of the proposed methods.

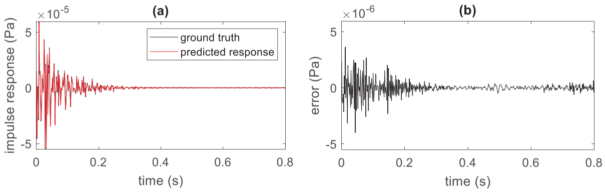

Figure 13. (a) Comparison of real and predicted IRs (b) Error curve.

Table 2. Errors.

	Absolute error	Percentage error
T_{60} (s)	0.02	3.50%
C_{50} (dB)	0.23	4.93%
C_{80} (dB)	0.25	2.68%
E_∞ (dB)	0.15	0.30%

Compared with the interpolation method, whereby the IRs are stored in grid with a spacing of one-fifth of the wavelength, the proposed method contributes to an improvement in storage reduction, whereby the storage of NNs is only 16.50 MB which is 4.29% of the interpolation method. The mean errors of 100 samples on 60 dB reverberation time T_{60}, 50 ms Clarity C_{50}, 80 ms clarity C_{80}, and total energy E_∞ are given in Table 2. The percentage errors are no more than 5%, which can't be noticed according to the just noticeable difference (JND).

4. CONCLUSIONS AND DISCUSSIONS

This project proposed a frequency-domain-based low-frequency room sound field modeling method for real-time auralization using a CNN trained by wave-based calculations. This method is applicable to the scenario of sound fields with moving sources and receivers, and varying head orientation. At the cost of training NNs, it contributes to obtaining relatively accurate BRIRs with the advantages of (i) real-time calculation: The output of NNs are SH coefficients of sound fields. The plane-wave decomposition is avoided at the run-time and the SH coefficients of sound fields can be combined with the SH coefficients of HRTFs to compute BRIRs efficiently. (ii) storage data reduction: Only NNs and SH coefficients of HRTFs should be stored to realize fast auralization. This approach is attractive to be used in the enhancement of Virtual Reality systems.

The future of our work will be enriched with (i) further application of the method in 3D scenes; (ii) training the NNs with varying wall impedances and room geometries to predict sound fields in rooms with different materials and dimensions; (iii) using non-uniform sampling methods to reduce the large pre-calculation requirements of FEM; and (iv) combining high-frequency method, applying HRTFs to obtain the BRIRs, and conducting listening tests.

5. REFERENCES

[1] F. Pind, C. Jeong, H. S. Llopis, K. Kosikowski, and J. Strømann-Andersen: “Acoustic Virtual Reality - Methods and challenges.” Baltic-Nordic Acoustic Meeting, 2018.

[2] W. Wittebol, and M. C. J. Hornikx: “A hybrid room acoustic approach for auralization.” Euronoise, 2021.

[3] J. Sheaffer, M. V. Walstijn, B. Rafaely, and K. Kowalczyk: “Binaural reproduction of finite difference simulations using spherical array processing.” IEEE/ACM Transactions on Audio Speech & Language Processing, 23(12), pp. 2125-2135, 2015.

[4] R. Mehra, L. Antani, S. Kim and, D. Manocha: "Source and Listener Directivity for Interactive Wave-Based Sound Propagation," in IEEE Transactions on Visualization and Computer Graphics, 20(4), pp. 495-503, 2014.

[5] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, C. A. Deledalle, and W. Li: “Machine learning in acoustics: a review.” J. Acoust. Soc. Am. 2019.

[6] V. Pulkki, and U. P. Svensson: “Machine-learning-based estimation and rendering of scattering in virtual reality.” J. Acoust. Soc. Am. 145(4), pp. 2664-2676, 2019.

[7] R. A. Tenenbaum, F. O. Taminato, and V. Melo: “Fast auralization using radial basis functions type of artificial neural network techniques.” Appl. Acoust. 157, 106993, 2020.

[8] E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstof: “Generative models for sound field reconstruction.” J. Acoust. Soc. Am. 153, pp. 1179-1190, 2023.

[9] N. Borrel-Jensen, A. P. Engsig-Karup, and C. H. Jeong: “Physics-Informed Neural Networks (PINNs) for Sound Field Predictions with Parameterized Sources and Impedance Boundaries.” JASA Express Lett. 12(1), 122402, 2021.

[10] A. Ratnarajah, Z. Tang, R. C. Aralikatti, and D. Manocha: "Neural Acoustic Impulse Response Generator for Complex 3D Scenes." https://doi.org/10.48550/arXiv.2205.09248

[11] A. Luo, Y. Du, M. J. Tarr, J. B. Tenenbaum, A. Torralba, and C. Gan: "Learning Neural Acoustic Fields." https://doi.org/10.48550/arXiv.2204.00628

[12] H. Gamper, and I. J. Tashev: "Blind reverberation time estimation using a convolutional neural network." International Workshop on Acoustic Signal Enhancement, 2018.

[13] A. F. Genovese, H. Gamper, V. Pulkki, N.Raghuvanshi, and I. J. Tashev: "Blind Room Volume Estimation from Single-channel Noisy Speech." IEEE International Conference on Acoustic, Speech and Signal Processing, 2019.

[14] C. Foy, A. Deleforge, and D. D. Carlo: "Mean absorption estimation from room impulse responses using virtually supervised learning." J. Acoust. Soc. Am. 150, pp. 1386-1299, 2021.

[15] O. Lundin, E. Zea, J. Cuenca, and U. P. Svensson: "Prediction of eigenfrequencies in non-rectangular rooms with machine learning." 24th International Congress on Acoustics, 2022.

[16] G. Götz, R. F. Pérez, S. J. Schlecht, and V. Pulkki: "Neural network for multi-exponential sound energy decay analysis." J. Acoust. Soc. Am. 52, pp. 942-953, 2022.

[17] M. Pollow, K. Nguyen, O. Warusfel, T. Carpentier, M. Müller-Trapet, M. Vorländer, and M. Noisternig: "Calculation of Head-Related Transfer Functions for Arbitrary Field Points Using Spherical Harmonics Decomposition." Acta Acustica united with Acustica, 98(1), pp. 72-82, 2012.

[18] B. Rafaely: Fundamentals of spherical array processing. Springer: Berlin Heidelberg, pp. 57-83, 2015.

[19] E. G. Williams: Fourier Acoustics: Sound Radiation and Near-Field Acoustical Holography. New York: Academic, pp. 217-221, 1999.

[20] https://nl.mathworks.com/help/deeplearning/ug/train-a-variational-autoencoder-vae-to-generate-images.html