DOI: 10.61782/fa.2023.0542

forum acusticum 2023

ROOM DIMENSION AND BOUNDARY CONDITION INFERENCE USING ROOM TRANSFER FUNCTIONS

Yuanxin Xia[1] **Allan P. Engsig-Karup**[2] **Cheol-Ho Jeong**[1]*

[1] Acoustic Technology, Department of Electrical and Photonics Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
[2] Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

ABSTRACT

The estimation of the absorption coefficients of the boundary surfaces in a room is important in room acoustic engineering. This research presents a machine learning method learns from simulated data to estimate the room dimensions and frequency-dependent absorption coefficients. We employ multi-task convolutional neural networks for inferring the frequency-dependent absorption coefficients and the dimensions of the room from transfer functions calculated by wave-based room acoustic methods. The proposed method provides reasonably accurate estimation of the boundary conditions and dimensions.

Keywords: Machine learning, absorption coefficient, room dimension, room transfer functions.

1. INTRODUCTION

Humans cannot estimate a room's dimensions and sound absorption configuration via hearing only. Sound travels so fast, and therefore the reflection overlap is heavy in rooms even within a short time from a sound generation. Knowing the sound absorption distribution is not considered important on a daily basis, but it gets more important in practical room acoustic engineering works. For example, acousticians would need to simulate several absorption configurations to find the optimum treatment of absorbers and scatters in a refurbishment/intervention project.

There are several methods proposed to inversely estimate the room geometry from room impulse responses (RIR) [1]. The main idea is to use lower-order specular reflections to estimate the times and directions of arrivals. Utilizing multiple source-receiver pairs makes it easier to estimate the geometry of a room [2, 3]. Generally, these analyses are conducted in the time domain mainly using simulated reflectograms from geometrical acoustics simulations, e.g., the image source method, where waves are simplified as a bundle of straight rays. Therefore, wave phenomena, such as diffraction and interference, are often neglected in such data.

In contrast, a transfer function (TF) in the frequency domain, the Laplace transform pair of a corresponding RIR, includes precise information about the amount of absorption via two different ways: a frequency shift from the theoretical natural frequency and the broadness of the peaks. The primary focus of this paper is to estimate the low to mid-frequency absorption configuration and extract the room dimensions from TFs via machine learning approaches.

2. DATASET GENERATION

The absorption coefficients in room acoustic engineering are typically necessary up to 4 kHz according to ISO 3382-1 [4]. However, due to computational limitations, we limited the frequency range for the TF dataset to

*Corresponding author: chje@dtu.dk.

250 Hz octave band with 0.5 Hz frequency resolution. The size-corrected random-incident absorption coefficients of each surface over three-octave bands, 63 Hz, 125 Hz, and 250Hz, as well as the length, width, and height are chosen as a label series for one room transfer function.

The Latin hypercube sampling is used to randomize the sources and receivers in a room to ensure a well-distributed sampling of the room's transfer function as shown in Figure 2, we use 6 sources corresponding to 125 receivers. We used two types of porous materials. A: Ecophon Akusto Wall-A (40 mm glass wool with a specific flow resistivity of 47,000 $\mathrm{Ns/m^4}$) and B: Ecophon Industry Modus (100 mm glass wool with a specific flow resistivity of 10,900 $\mathrm{Ns/m^4}$). The absorption coefficient of the remaining concrete surfaces is set to 0.029, 0.048, 0.043 for the 63 Hz, 125 Hz, and 250 Hz octave bands, respectively. In Figure 1, a rectangular room is shown with numerical labels assigned to the surfaces. To generate the training data, material configurations are varied according to Table 1, with a dash indicating the concrete surface. The room size variation is shown in Table 2, following guidelines for low-frequency optimization, as detailed in Ref. [5]. Each configuration corresponds to 7 room sizes, 6 sources, and 125 receivers, resulting in 5,250 TFs.

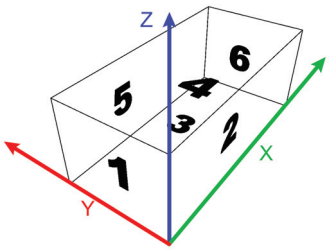

Figure 1: Surface number assignment.

The dataset is generated by the commercial software COMSOL® Multiphysics with MATLAB® LiveLink™ by a single Intel Xeon Gold 6226R CPU. Dataset augmentation is a standard practice in machine learning. In this study, we utilized loudspeaker frequency response to augment a simulated dataset, resulting in two datasets: "Simu." (pure) and "Aug." (augmented). This approach is expected to enhance the performance machine learning models.

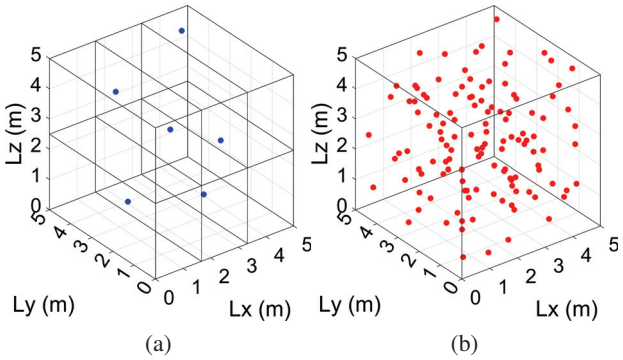

Figure 2: Example of the distribution of sources (a) and receivers (b) in a room.

3. NEURAL NETWORKS STRUCTURE AND TRAINING STRATEGY

In this paper, we employ the standard ResNet V2 architecture as our backbone model, as outlined in Ref. [6], due to its exceptional performance in optimizing the loss function [7]. As noted in Ref. [8], increasing the depth of neural networks can reduce test error; however, the error begins to rise beyond a specific threshold. Therefore, it is essential to align the network's depth with the problem's complexity, therefore, we decided to use ResNet18.

The activation function for the absorption coefficient outputs in each layer is chosen as the sigmoid function, providing results in the range of zero to one, aligning with the definition of the absorption coefficient. The activation function for the room dimension output layer is the rectified linear unit (ReLU), resulting in outputs between zero and infinity. The loss function is comprised of two components: the absorption coefficient and the room dimension, as detailed below:

$$\mathcal{L}(\theta) = \lambda_d \sum_{i=1}^{N_d} (\boldsymbol{y}_d^{(i)} - f_{\boldsymbol{\theta}_d}(\mathbf{x}))^2 + \sum_{j=1}^{N_f \cdot N_s} \boldsymbol{\lambda_a}^{(j)} (\boldsymbol{y}_a^{(j)} - f_{\boldsymbol{\theta}_a}(\mathbf{x}))^2. \tag{1}$$

Here, λ_d represents the weight assigned to the room dimension estimation task, while $\boldsymbol{\lambda_a}$ signifies the weight of the absorption coefficient for each frequency band and surface. N_f and N_s denote the number of frequency bands and surfaces, respectively. $\mathbf{x}$ corresponds to the input TF, while $\boldsymbol{y}_d$ and $\boldsymbol{y}_a$ indicate the associated room dimension and absorption coefficient labels.

Table 1: Material assignment of each surface.

Surface	1	2	3	4	5	6	Surface	1	2	3	4	5	6
Config 1	**A**	-	-	-	-	-	Config 15	-	**A**	-	-	**A**	-
Config 2	**B**	-	-	-	-	-	Config 16	-	**B**	-	-	**B**	-
Config 3	-	**A**	-	-	-	-	Config 17	**A**	-	-	-	-	**A**
Config 4	-	**B**	-	-	-	-	Config 18	**B**	-	-	-	-	**B**
Config 5	-	-	**A**	-	-	-	Config 19	**A**	-	-	**A**	-	**A**
Config 6	-	-	**B**	-	-	-	Config 20	**B**	-	-	**B**	-	**B**
Config 7	-	-	-	**A**	-	-	Config 21	**A**	-	-	**B**	-	**A**
Config 8	-	-	-	**B**	-	-	Config 22	**B**	-	-	**A**	-	**B**
Config 9	-	-	-	-	**A**	-	Config 23	-	**A**	**A**	-	**A**	-
Config 10	-	-	-	-	**B**	-	Config 24	-	**B**	**B**	-	**B**	-
Config 11	-	-	-	-	-	**A**	Config 25	-	**A**	**B**	-	**A**	-
Config 12	-	-	-	-	-	**B**	Config 26	-	**B**	**A**	-	**B**	-
Config 13	-	-	**A**	**A**	-	-	Rigid	-	-	-	-	-	-
Config 14	-	-	**B**	**B**	-	-	Summary	8	8	8	8	8	8

Table 2: Aspect ratio of the room.

Room Ratio	Length(m)	Width (m)	Height (m)	Area (m^2)	Volume (m^3)
1:1.11:1.67	3.0	4.5	2.7	13.50	36.45
∼Bolt (2:3:5)	4.0	6.75	2.7	27.00	72.90
∼Louden (1:1.4:1.9)	3.8	5.15	2.7	19.47	52.84
∼Cox(1:1.56:1.86)	4.2	5.0	2.7	21.00	56.70
1:1.33:2.66	4.0	8.0	3.0	32.00	96.00
1:1.67:1.67	5.0	5.0	3.0	25.00	75.00
1:1:1.3	4.3	3.3	3.3	14.19	46.83

In this research, we construct the neural network using the TensorFlow framework (Figure 3). We applied training parameters such as an 80:20 split for training and testing data, a batch size of 32, and 200 epochs. The Adam optimization algorithm was employed with an initial learning rate of 1×10^{-4}. We set the number of filters in the input layer to 64, while maintaining the remaining network structure according to the residual network principles. To prevent overfitting, we implemented batch normalization post-convolution [9] and utilized Kaiming initialization for weight initialization [10]. All experiments were conducted using the NVIDIA Tesla V100 GPU.

Determining loss weights in Multi-task learning (MTL) remains an active area of investigation. It is crucial to acknowledge that the overall loss is dominated by the small gradient term in MTL. Therefore, more challenging tasks should be assigned with larger loss weights to balance the overall loss. In this study, we considered room dimension estimation a simpler task than absorption coefficient estimation, a more abstract feature of the room transfer function and consequently more challenging to predict. We also observed that predicting the absorption coefficient became more difficult with increasing frequency. In our initial experiment, we allocated a uniform loss weight to all tasks and removed the room dimension branch to avoid potential disturbances.

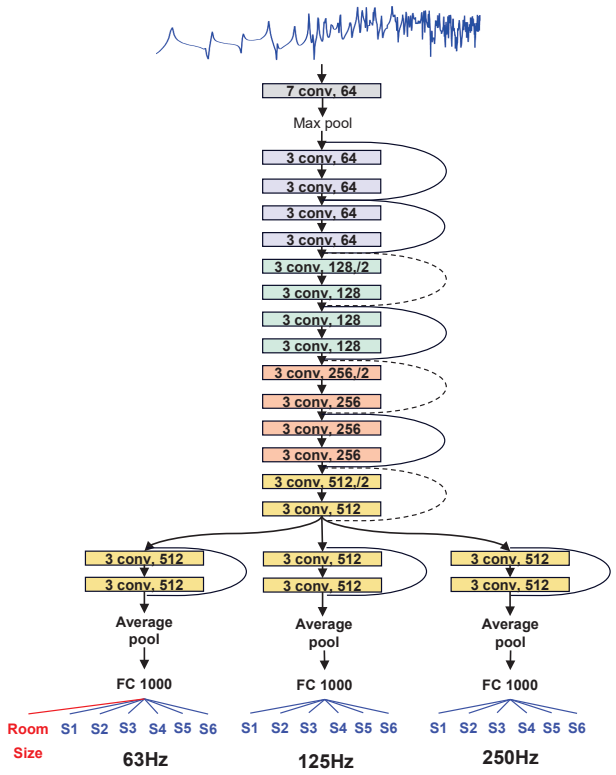

Figure 3: Multi-task residual networks for absorption coefficient and room dimension estimation.

4. RESULTS AND DISCUSSION

In this section, we assess the performance of the networks on test sets, monitoring the loss associated with two categories of tasks under varying training strategies. This loss quantifies the discrepancy between the network's predictions and the actual ground truth labels, where a lower loss indicates a good prediction of the network.

As depicted in Figure 4, the loss of the absorption coefficient increases almost linearly with the frequency bandwidth, in accordance with the constant percentage bandwidth rule. The neural networks aim to provide equally precise predictions for all absorption coefficients regardless of the frequency band. Therefore, it is important to balance the absorption loss by prioritizing difficult tasks. One approach is to adjust the weights for the absorption coefficient loss proportionally to the frequency bandwidth. Specifically, the weight of the neighboring higher-frequency octave band is twice as high as that of the current octave band, different from the uniform distribution, this strategy is named as frequency-dependent weights (FDW) in Figure 5. The loss metric is calculated as $\Delta\alpha = \frac{1}{N_f \cdot N_s} \sum_{i=1}^{N_s} \sum_{j=1}^{N_f} (\alpha_{i,j}^{est} - \alpha_{i,j}^{true})^2$. The application of the FBW method results in a reduction of the loss metric. Furthermore, the mixed dataset (Simu. and Aug.) outperforms the single dataset.

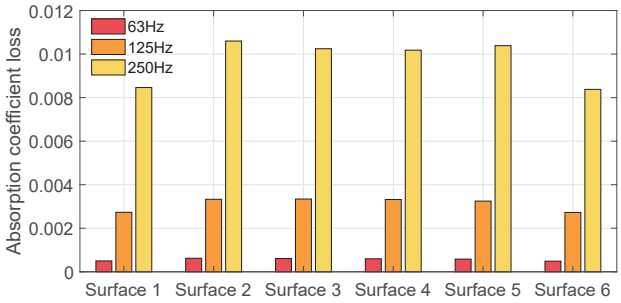

Figure 4: Absorption coefficient loss as a function of frequency for the test dataset.

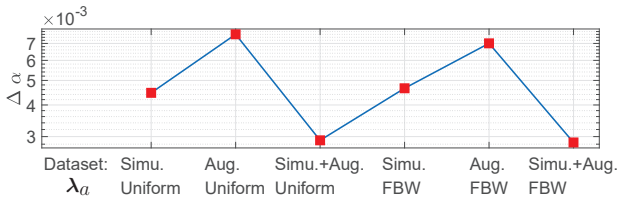

Figure 5: Comparison of training strategies of absorption coefficient branches.

Employing the optimal training strategy of absorption coefficient branches, we further examine the impact of room dimension branch weight λ_d in reference to the sum of the individual weights of $\boldsymbol{\lambda}_a$. The results of this analysis are illustrated in Figure 6. The loss metric is defined as $\Delta L = \frac{1}{N_d} \sum_{i=1}^{N_d} (L_i^{est} - L_i^{true})^2$. While a small value of λ_d may provide a reasonably precise estimation of the room dimensions, an increased λ_d emphasizes the room dimension regression task, thereby yielding a more accurate estimation.

5. CONCLUSION

This study evaluates the feasibility of a ResNet architecture to estimate the room dimensions and frequency-independent absorption coefficients, using measured

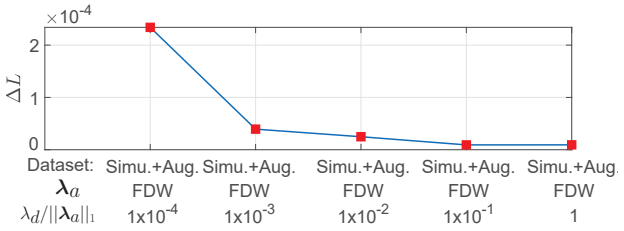

Figure 6: Impact of room dimension branch weight λ_d.

transfer functions up to the 250 Hz octave band. The ResNet18 shows promising results in estimating the dominant directions for the energy decay, but precise estimations of the absorption coefficients of a parallel surface pair were not guaranteed. In this study, the training data generation took 99.66% (approximately 40 hours per configuration), and the model training took 0.34% of time (approximately 4 hours). Once the model has been trained and loaded, one inference takes around 20 ms.

Moving forward, our intention is to enhance the model's resilience by incorporating time domain information and physics constraints, thereby reducing reliance solely on frequency domain information. Our ultimate goal is to apply this refined method to actual measurement data. Though we may encounter unpredictable variables inherent in real-world scenarios, such challenges provide valuable opportunities for improvement.

6. REFERENCES

[1] I. Dokmanić, Y. M. Lu, and M. Vetterli, "Can one hear the shape of a room: The 2-d polygonal case," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 321–324, IEEE, 2011.

[2] S. Tervo and T. Tossavainen, "3d room geometry estimation from measured impulse responses," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 513–516, IEEE, 2012.

[3] S. Park and J.-W. Choi, "Iterative echo labeling algorithm with convex hull expansion for room geometry estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1463–1478, 2021.

[4] ISO, "Iso 3382-1:2009 acoustics – measurement of room acoustic parameters – part 1: Performance spaces," standard, International Organization for Standardization, Geneva, Switzerland, 2009.

[5] T. J. Cox, P. D'Antonio, and M. R. Avis, "Room sizing and optimization at low frequencies," *Journal of the Audio Engineering Society*, vol. 52, no. 6, pp. 640–651, 2004.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*, pp. 630–645, Springer, 2016.

[7] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," *Advances in neural information processing systems*, vol. 31, 2018.

[8] E. Nichani, A. Radhakrishnan, and C. Uhler, "Increasing depth leads to u-shaped test risk in over-parameterized convolutional networks," *arXiv preprint arXiv:2010.09610*, 2020.

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.