# ROOM IMPULSE RESPONSE PREDICTION WITH NEURAL NETWORKS: FROM ENERGY DECAY CURVES TO PERCEPTUAL VALIDATION

*Imran Muhammad, Gerald Schuller*

Technische Universität Ilmenau, Germany

## ABSTRACT

Prediction of room impulse responses (RIRs) is essential for room acoustics, spatial audio, and immersive applications, yet conventional simulations and measurements remain computationally expensive and time-consuming. This work proposes a neural network framework that predicts energy decay curves (EDCs) from room dimensions, material absorption coefficients, and source–receiver positions, and reconstructs corresponding RIRs via reverse-differentiation. A large training dataset was generated using room acoustic simulations with realistic geometries, frequency-dependent absorption, and diverse source–receiver configurations. Objective evaluation employed root mean squared error (RMSE) and a custom loss for EDCs, as well as correlation, mean squared error (MSE), spectral similarity for reconstructed RIRs. Perceptual validation through a MUSHRA listening test confirmed no significant perceptual differences between predicted and reference RIRs. The results demonstrate that the proposed framework provides accurate and perceptually reliable RIR predictions, offering a scalable solution for practical acoustic modeling and audio rendering applications.

***Index Terms***— Room Acoustics, Neural Network, Energy Decay Curves, Auralization, Virtual Reality

## 1. INTRODUCTION

The room impulse response plays a vital role in room acoustics, spatial audio rendering, and immersive virtual environments. Standard room acoustic parameters, which provide quantitative insight into acoustic behavior of the closed space, are often derived from RIRs or EDCs, which describe how sound energy evolves over time after an acoustic excitation [1]. Classical room-acoustic modeling relies on computationally expensive simulations (e.g., ray tracing (RT), image-source method (ISM), and wave-based simulations) [2] or time-consuming on-site measurements, which limit scalability and flexibility. Fast acoustics software such as pysound [3], ITA Geometrical Acoustics [4] openRay [5], EVERTims [6], GSoundSIR [7] and Pyroomacoustics [8] enable an efficient alternative to physical measurements. However, these methods are based on approximations and struggle to capture wave-based phenomena accurately, particularly at low frequencies. Wave-based solvers offer greater accuracy but are often computationally intensive and unsuitable for real-time applications. This trade-off between speed and accuracy presents an opportunity for deep learning or data-driven approaches. Crucially, such models should be trained on high-quality data, ideally derived from physical measurements or wave-based simulations, to ensure reliability.

Data-driven approaches have grown rapidly in the last few years, targeting either (i) prediction of room parameters from descriptors of the space, [9] (ii) direct estimation or completion of room impulse responses (RIRs), [10, 11] or (iii) downstream dereverberation and robust ASR. Early neural predictors focused on mapping geometric and material features to scalar reverberation metrics (e.g., T60, clarity), demonstrating that learned regressors can approximate or outperform simplified formula-based estimates in complex rooms [12, 13, 14, 15]. Beyond scalar metrics, several works aim to reconstruct full RIRs. Deep RIR completion formulates RIR completion as a lightweight learning task: given partial or bandlimited information, a neural model infers broadband RIRs, evaluating with EDF/EDC errors, T60 MSE, and DRR MSE across multiple datasets [16, 17, 10]. These advancements highlight the effectiveness of deep learning in capturing the intricate temporal and spatial characteristics of room acoustics, providing efficient and scalable alternatives to conventional physics-based modeling techniques.

We propose a neural network–based framework for estimating RIRs by first predicting EDCs directly from input parameters describing the acoustic environment. The input features include room dimensions, material absorption properties, and source–receiver positions, which collectively define the acoustical behavior of a given space. In contrast, RIRs are high-dimensional signals with complex temporal and spectral details, making direct prediction more challenging and error-prone. Since EDCs also serve as the basis for computing important acoustic parameters [2], this strategy simplifies the modeling task while preserving the essential information required for an accurate room acoustics analysis. This framework centers on a neural model that maps room geometry and material properties to EDCs. Using EDCs as an intermediate representation captures temporal energy decay, which are then used to reconstruct full RIRs through reverse differentiation and polarity assignment. This ensures

consistency of temporal and spectral features while leveraging the stability of EDC modeling. Training relied on a large simulated dataset incorporating varied geometries, frequency-dependent absorption, and source–receiver configurations to enable generalization. Evaluation combined objective and perceptual measures: root mean square error (RMSE) and a custom temporal-decay loss for EDCs, whereas correlation, RIR-MSE, and spectral MSE for RIRs provided a comprehensive assessment of accuracy, while perceptually motivated approaches such as psychoacoustic losses have also been explored in related work [18].

Finally, perceptual validation was conducted through a MUSHRA listening test. The results demonstrated that listeners perceived no significant differences between the predicted and reference RIR conditions, highlighting the effectiveness of the proposed method in delivering perceptually convincing outcomes. This confirms that the framework not only achieves strong numerical performance but also satisfies perceptual standards, making it a promising tool for applications in virtual acoustics, audio rendering, and acoustic design.

## 2. METHODOLOGY

We propose a deep learning framework based on Long Short-Term Memory (LSTM) neural networks to predict energy decay curves (EDCs) directly from room geometries and wall absorption characteristics. LSTMs are well suited for sequential modeling as they capture long-term temporal dependencies and mitigate vanishing gradients, a limitation of standard recurrent neural networks (RNNs) [19, 20]. Compared to convolutional neural networks (CNNs), which focus on spatial pattern recognition, LSTMs are more effective for modeling the temporal decay of EDCs. While Transformer architectures achieve strong sequence modeling performance, their computational demands and data requirements make them less practical in this setting [21]. LSTMs therefore offer a robust and efficient solution for predicting EDCs from room features. The model is trained on synthetic dataset covering varied room dimensions, source–receiver placements, and surface absorption properties. Input features are normalized and structured for LSTM model, and training is guided by a custom loss function (Sec. 2.2) tailored for accurate EDC prediction.

### 2.1. Dataset and Room Parameters

A dataset of 6000 simulated shoebox rooms was generated using the Pyroomacoustics library [22], with each room defined by length (L), width (L), height (H), source/receiver 3D positions (X, Y, Z), and frequency-dependent absorption coefficients averaged across surfaces [8]. Room dimensions, positions, and material properties were systematically varied (Table 1) to span realistic acoustic conditions. Pyroomacoustics combines the image source method (ISM) and ray tracing to approximate high-frequency sound propagation, efficiently generating RIRs and EDCs at $48\,\mathrm{kHz}$. While limited by its geometric assumptions (ignoring wave effects), this approach covers a broad range of reverberation times (T60), from nearly anechoic ($0.1\,\mathrm{s}$) to highly reverberant ($3.0\,\mathrm{s}$).

**Table 1**. Range of simulated room features used as input parameters for LSTM model training

| **Parameter** | **Range / Values** |
|---|---|
| Room Dimensions (L×W×H) | 3-6 m × 3-6 m × 2.5-4 m |
| Source-Receiver Distances | 1 - 4 m |
| Wall Absorptions Range | 0.14 - 0.65 (125 - 8000 Hz) |
| No. of Room Configurations | 6000 |
| No. Material Types | 10 |

Input features were scaled using *MinMax* normalization to $[0, 1]$ for consistent training dynamics. Although EDCs already fall within this range, normalization was applied as a standard deep learning practice. Data were reshaped into $[N, 1, 16]$, where $N$ is the number of rooms and 16 the features per room, to match the input format required by the LSTM architecture (batch_size, sequence_length, feature_dimension). Each room sample is thus represented as a one-step sequence of 16 values, enabling the LSTM to capture potential inter-dependencies. For evaluation, the dataset was split into $60\%$ training, $20\%$ validation, and $20\%$ testing.

### 2.2. LSTM Model Architecture

The LSTM architecture was implemented in PyTorch with 128 hidden units to capture dependencies among input features via memory cells and gating mechanisms. To reduce overfitting, a $30\%$ dropout was applied after the LSTM. This was followed by a dense layer of 2048 *ReLU*-activated neurons for non-linear feature learning, with another dropout for regularization. The output layer matched the EDC sequence length, using a linear activation for continuous prediction. Training used the Adam optimizer ($LR = 0.001$) and mean squared error (MSE) with an additional custom loss. Early stopping (patience of 10) prevented overfitting and improved efficiency. Although training was capped at 200 epochs, convergence often required fewer iterations.

We designed a composite loss function that integrates two complementary terms: EDCs fidelity, RIR fidelity. Let $\mathbf{y} \in \mathbf{R}^T$ denote the ground-truth (target) EDC of length $T$, and $\hat{\mathbf{y}} \in \mathbf{R}^T$ the predicted EDC. The corresponding RIR signals are obtained by using RSS approach in Section 2.3. The total loss is defined as a weighted sum of two terms.

$$\mathcal{L} = \alpha \left(\frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_t - y_t\right)^2\right) + \beta \left(\frac{1}{T-1}\sum_{t=2}^{T}\left(\hat{r}_t - r_t\right)^2\right) \quad (1)$$

Where $\alpha, \beta \geq 0$ are weighting coefficients and $\mathbf{r}_t$ denotes

the RIR obtained from the ground-truth EDC, and $\hat{\mathbf{r}}t$ denotes the RIR obtained from the predicted EDC. We choose $\alpha = 1$ and $\beta = 0.5$. In the Eq. 1, the EDC loss enforces accurate cumulative decay prediction, where as the RIR loss constrains the predicted EDC to yield a realistic RIR shape.

### 2.3. Reconstruction of Room Impulse Responses from Energy Decay Curves

For a given a room impulse response $h(n)$, its instantaneous energy is defined as

$$e(n) = h^2(n), \tag{2}$$

and the cumulative decay curve (EDC) is computed in reverse time as

$$\mathrm{EDC}[n] = \sum_{k=n}^{N-1} e[k], \tag{3}$$

where $T$ is the signal length. Differentiating the EDC redistributes the energy over time:

$$d[n] = -\big(\mathrm{EDC}[n+1] - \mathrm{EDC}[n]\big), \tag{4}$$

and the magnitude response of the RIR can be estimated as

$$h_{\mathrm{mag}}[n] = \sqrt{\max(d[n], 0)} \tag{5}$$

Since only the magnitude can be recovered from the EDC, the polarity/phase must be reconstructed using stochastic methods. Two strategies are considered: i) the *random sign* method, where each sample is assigned an independent random polarity, and ii) the *random sign-sticky* method, which enforces sign continuity between consecutive samples. In our implementation, continuity is controlled by a *stickiness parameter*, set to $0.90$, meaning that each sample has a $90\%$ probability of retaining the polarity of the previous one and a $10\%$ chance of flipping. A higher stickiness value promotes smoother, more coherent low-frequency oscillations, which is beneficial for long-reverberant RIRs, while a lower value allows faster sign changes, enhancing the representation of short reverberation tails. Empirically, stickiness values in the range $0.7$–$0.95$ were found to yield stable and perceptually plausible reconstructions across a variety of reverberation times.

## 3. EVALUATION

After training, the model demonstrates strong predictive accuracy across unseen room configurations. The accuracy of the predicted EDCs are evaluated to verify the accuracy of the model. The predicted EDCs are then used to compute the room acoustical parameters such as reverberation time (T20), early decay time (EDT) and clarity (C50). These parameters are compared with those derived from the target EDCs to further access the reliability of the model.

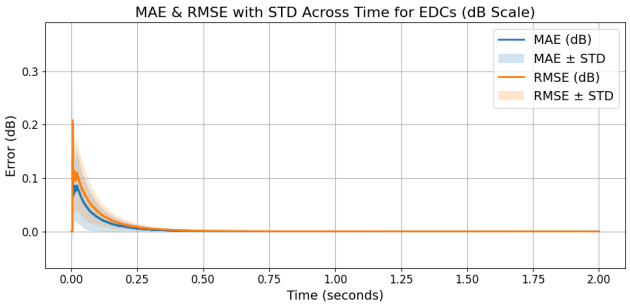


**Fig. 1**. **MAE** and **RMSE** of the predicted EDCs, averaged over time across all room configurations

### 3.1. Objective Evaluation of EDCs

To assess the model's predictive performance, qualitative comparisons were conducted between the target (i.e. simulated) and predicted EDCs. We reported the MAE and RMSE computed across all predicted EDCs, averaged over time. Additionally, we included the standard deviation to reflect variability across the dataset as shown in Figure 1 to provide a more comprehensive evaluation of model behavior. The graph reveal a strong agreement between the predicted and target EDC curves, suggesting that the model effectively captures the temporal decay characteristics of rooms. The predicted EDCs not only replicate the overall decay envelope but also capture fine-grained fluctuations in the decay profile.

### 3.2. Objective Evaluation of Reconstructed RIRs

To objectively assess the quality of the reconstructed room impulse responses from the predicted energy decay curves, we computed several signal-level metrics, including the MSE between actual and predicted RIRs, correlation coefficients, and spectral MSE. RIRs were reconstructed from EDCs using the RSS approach. The objective results indicated that the reconstructed RIRs closely approximate the temporal and spectral characteristics of the ground-truth RIRs. From the four randomly selected predicted EDCs, the corresponding constructed RIRs showed stronger correlations ($0.37 - 0.52$) and lower spectral MSE ($55$ - $58\,\mathrm{dB}$). Table 3 summarizes the error metrics. These objective findings are corroborated by the perceptual evaluation conducted via a MUSHRA listening test, where the RSS-based reconstructed RIRs were rated significantly closer to the reference RIRs (sec 3.4). Overall, the analysis demonstrates that RIRs can be effectively reconstructed from predicted EDCs using the Random Sign-Sticky method, achieving both objective similarity and perceptual plausibility.

### 3.3. Acoustic Parameter Estimation

Additional analyses such as key room acoustic parameters (T20, EDT and C50) were derived from both the target and predicted EDCs. The predictive performance was quantified

**Table 2**. Objective comparison of the four reconstructed RIRs using the RSS-approach. Metrics include RIR mean squared error (RIR-MSE), correlation coefficient, and spectral MSE.

| No. | RIR-MSE | **Correlation** | **Spectral MSE (dB)** |
|---|---|---|---|
| 1 | $1.20 \times 10^{-5}$ | 0.515 | 58.26 |
| 2 | $1.22 \times 10^{-5}$ | 0.493 | 57.26 |
| 3 | $1.27 \times 10^{-5}$ | 0.431 | 54.98 |
| 4 | $1.45 \times 10^{-5}$ | 0.369 | 54.64 |

**Table 3**. Performance of LSTM Model: MAE, RMSE, and coefficient of determination ($R^2$) computed between predicted and target EDCs

| **Metric** | **MAE** | **RMSE** | **$R^2$** |
|---|---|---|---|
| EDT (s) | 0.033 | 0.042 | 0.92 |
| T20 (s) | 0.023 | 0.029 | 0.97 |
| C50 (dB) | 0.91 | 2.53 | 0.84 |

using MAE, RMSE, and $R^2$ (coefficient of determination) for each parameter across all room configurations in the test dataset. Table 3 summarizes the error metrics. The model shows a good performance in estimating EDT with a MAE of $0.033$ s and an RMSE of $0.042$ s and T20 with MAE of $0.023$ s and an RMSE of $0.029$ s respectively. The high $R^2$ value of $0.92$ and $0.97$ for EDT and T20 respectively indicates that the predictions closely follow the trend of target values, confirming the model's ability to learn the early decay slope of the EDCs.

### 3.4. Perceptual Validation of RIR Reconstruction

We evaluated the perceptual plausibility of reconstructed RIRs generated from predicted EDC. From these predicted EDCs, RIRs were generated and subsequently compared against RIRs obtained from actual EDCs, which served as the perceptual reference. The reconstructed RIRs from both methods in section 2.3, were used to convolve four audios (*Casta (music)*, *Mix (Music+Speech)*, *Male Speech*, and *o444 (voice-based)*). Four different RIRs were selected with reverberation times ($T20 = 0.5s, 1.34s, 2.0s$, and $3.0s$). For each audio-reverberation combination, four stimulus conditions were generated: the *reference* (RIR synthesized from the actual EDC), one anchors (*Anchor35* [23], *Random-Sign* (RIR synthesized from the target EDC using RS approach) and *Random-Sign-Sticky* (RIR synthesized from the target EDC using RSS approach). This resulted in sixteen trials per participant, each trial containing four stimuli.

A MUSHRA listening experiment was conducted following ITU-R BS.1534 recommendations. We used webmushra software [23]. Twelve normal-hearing participants, all with prior experience in audio research or perceptual testing, rated the *basic audio similarity* of each stimulus on a $0-100$ scale,

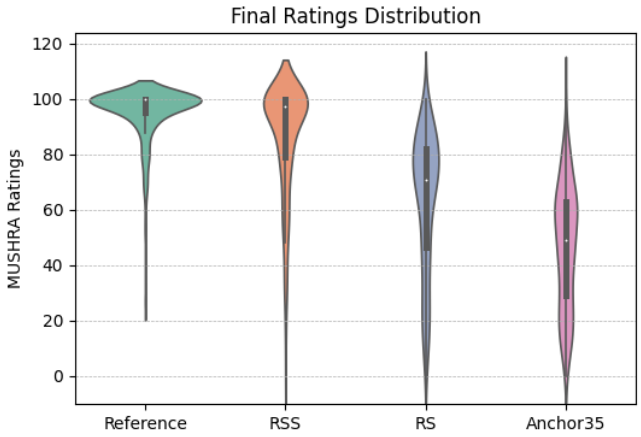


**Fig. 2**. MUSHRA rating score: Refence, RSS (random sign sticky), RS (random sign) and Anchor35

where $100$ corresponded to the hidden reference and $0$ to the worst quality. Trials were presented in randomized order, and listeners underwent a training session to familiarize themselves with the interface and the range of anchor degradations. The results are show in the Fig 2 and Table 4.

**Table 4**. Summary statistics for different MUSHRA rating stimuli

| Rating Stimulus | Mean | Std | Median | SEM | CI95 |
|---|---|---|---|---|---|
| Reference | 95.32 | 9.82 | 100.0 | 0.68 | 1.34 |
| RSS | 85.81 | 20.61 | 97.5 | 1.43 | 2.82 |
| RS | 63.17 | 24.93 | 71.0 | 1.73 | 3.41 |
| Anchor35 | 46.67 | 22.24 | 49.0 | 1.54 | 3.04 |

## 4. SUMMARY AND OUTLOOK

The proposed LSTM model predicts EDCs from room geometrical and material features, achieving low mean absolute error and root mean square error, and and high $R^2$ for EDT, T20 and C50 across unseen room configurations. RIRs reconstructed via the random sign–sticky method preserve temporal and spectral structure, including low-frequency coherence. These reconstructions were further validated through a MUSHRA listening experiment, showing mean ratings of $85.8$ for random sign-sticky compared to $95.3$ for the reference. The approach provides fast, scalable room acoustics prediction and auralization, suitable for interactive applications in VR and acoustic design. The source code, preprocessed dataset, and trained model used in this study are available at [24].

## 5. REFERENCES


[1] Heinrich Kuttruff, *Room Acoustics*, CRC Press, 5 edition, 2009.

[2] Michael Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, Springer, Cham, 2 edition, 2020.

[3] Carl Schissler and Dinesh Manocha, "Gsound: Interactive sound propagation for games," in *Audio Engineering Society Conference: 41st International Conference: Audio for Games*. Audio Engineering Society, 2011.

[4] Virtual Acoustics, "Itageometricalacoustics," 2025.

[5] EmergentMind, "openray," 2025.

[6] S. Laine, S. Siltanen, T. Lokki, and L. Savioja, "Accelerated beam tracing algorithm," *Applied Acoustics*, vol. 70, no. 1, pp. 172–181, 2009.

[7] Yongyi Zang and Qiuqiang Kong, "Gsound-sir: A spatial impulse response ray-tracing and high-order ambisonic auralization python toolkit," 2025.

[8] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 351–355.

[9] Cédric Foy, Antoine Deleforge, and Diego Di Carlo, "Mean absorption estimation from room impulse responses using virtually supervised learning," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2021.

[10] J. Lin, G. Götz, and S. J. Schlecht, "Deep room impulse response completion," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, no. 20, 2025.

[11] Xenofon Karakonstantis and et al., "Room impulse response reconstruction with physics-informed deep learning," *Journal of the Acoustical Society of America*, 2024.

[12] Chen Meng, Noam Shabtai, and Boaz Rafaely, "Predicting room acoustic parameters from room geometry using deep learning," *The Journal of the Acoustical Society of America*, vol. 154, no. 4, pp. 2452–2461, 2023.

[13] Karolina Prawda, Sebastian J. Schlecht, and Vesa Välimäki, "Calibrating the sabine and eyring formulas," *The Journal of the Acoustical Society of America*, vol. 152, no. 2, pp. 1158–1169, 08 2022.

[14] Georg Götz, Ricardo Falcón Pérez, Sebastian J. Schlecht, and Ville Pulkki, "Neural network for multi-exponential sound energy decay analysis," *The Journal of the Acoustical Society of America*, vol. 152, no. 2, pp. 942–953, 08 2022.

[15] Wangyang Yu and W. Bastiaan Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 436–447, 2021.

[16] Jihoon Kim and Youngmoo E Yang, "Generative adversarial neural network for room impulse response synthesis," *arXiv preprint arXiv:2311.02581*, 2023.

[17] Piotr Masztalski, Mateusz Matuszewski, Karol Piaskowski, and Michał Romaniuk, "Storir: Stochastic room impulse response generation for audio data augmentation," 2020.

[18] Gerald Schuller and Muhammad Imran, "A novel perceptual loss function for audio and music quality differentiable in pytorch," in *2024 58th Asilomar Conference on Signals, Systems, and Computers*, 2024, pp. 1418–1422.

[19] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] Yoshua Bengio, Patrice Simard, and Paolo Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[22] Robin Scheibler, Elie Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 351–355.

[23] Michael Schoeffler et al., "webmushra — a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, pp. 8, 2018.

[24] Gerald Schuller and Imran Muhammad, "Lstm-model-energy-decay-curves: Github repository (2025) link:," `https://github.com/TUIlmenauAMS/LSTM-Model-Energy-Decay-Curves`.