Audio Engineering Society

Conference Paper 26

Presented at the AES International Conference on Artificial Intelligence and Machine Learning for Audio
2025 September 8-10, London, UK

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

A Machine Learning approach to modal control in small rooms

Carlo M. Bolla[1], Trevor J. Cox[1], and Bruno M. Fazenda[1]

[1]Acoustic Research Centre, University of Salford, Salford M5 4WT, United Kingdom

Correspondence should be addressed to Carlo M. Bolla (c.m.bolla@edu.salford.ac.uk)

ABSTRACT

Reducing the decay time of low-frequency resonances in small rooms remains a critical challenge in audio engineering, as these resonances can degrade sound quality, leading to a "boomy" or "muddy" audio experience. Traditional equalization methods, while effective in certain contexts, rely heavily on static equalization filters, making them suboptimal for complex scenarios. This paper explores a novel machine learning-based approach to suppress low-frequency resonances by leveraging a modified Wave-U-Net architecture. The model incorporates psychoacoustic principles, specifically perceptual modal thresholds, to ensure that resonance suppression is guided by human auditory perception while preserving the natural characteristics of the audio. A large dataset of synthetic Room Impulse Responses was generated using a physically informed parametric model for cuboid rooms with randomized dimensions and absorption profiles. The model was trained on amplitude-modulated white noise bursts to simulate broadband excitation. Evaluation on unseen synthetic data demonstrates effective suppression of low-frequency resonances below perceptual thresholds. Limitations in generalization are discussed, along with future directions that include retraining on musical stimuli and incorporating perceptual loss metrics.

1 Introduction

Low-frequency resonances are a challenge in audio engineering, can impact the quality and clarity of audio in both recording and playback environments. These resonances, result in excessive buildup of energy at certain frequencies, leading to a "boomy" or "muddy" sound [1]. Addressing these is critical to ensuring that audio recordings and mixes translate well across different listening systems [2] [3] [4].

Conventional strategies for suppressing low-frequency room resonances have historically centred on the application of static equalisation filters, subsequently refined through expert manual adjustment to achieve the desired tonal balance.

Traditional approaches to managing low-frequency resonances typically involve the use of static equalization (EQ) filters [5]. These methods can be time-consuming, particularly in complex audio scenarios. Moreover, overuse of EQ can introduce unwanted artifacts, further compromising the audio quality [6]. Automated tools, such as dynamic EQs and resonance suppressors, have made strides in streamlining this process but often lack the precision needed to account for perceptual factors critical to listening experience [5]. Recent advancements in deep learning have demonstrated significant potential for using machine learning to address complex audio processing tasks.

Among these is Wave-U-Net, [7] a convolutional neural network architecture originally designed for audio source separation This has shown promise due to its ability to model time-frequency relationships effectively. It has been applied to tasks such as source separation [7], speech enhancement [8], and de-

reverberation [9]. However, its application to the low-frequency resonance suppression is yet to be fully explored.

This paper therefore presents an adaptation of the Wave-U-Net architecture to filter low-frequency resonances in audio. Our approach integrates psychoacoustic principles to align the filtering process with auditory perception.

By considering Perceptual Modal Thresholds (PMTs) [10] which define the threshold beyond which modal resonances become perceptually intrusive, we aim to suppress resonances to below perceptual significance rather than complete elimination.

Integrating these principles into deep learning models offers a promising avenue for perceptually-aware audio processing. Audio samples are provided online[1].

2 Method

A modified Wave-U-Net architecture was trained using supervised learning to reduce resonant room modes in reverberant signals to perceptually acceptable levels.

Primary focus was placed on verifying whether the proposed system could automatically detect modal frequencies whose decay exceeds the PMTs and adjust their decay accordingly, all within an end-to-end architecture.

While the model's non-causal design inherently precludes real-time processing, this was an exploratory phase, and other deep learning architectures suitable for real time processing will be tested later.

Given the impracticality of capturing real-world Room Impulse Responses (RIRs) across diverse acoustic conditions, we employed a parametric modal decomposition model derived from the formal solution of the wave equation [11] to synthesize a comprehensive training dataset.

Such a model, validated in other works [12] for accurately reproducing frequency responses of enclosed spaces within our target low-frequency range (20-240 Hz), enabled generation of approximately 2.7 million synthetic RIRs.

The dataset encompassed varied room dimensions, absorption characteristics, and randomized source-receiver placements (defined in Section 2.2), with careful constraints to avoid corner positions that could introduce bias through exaggerated modal effects.

The 240 Hz frequency cap was selected to focus computational resources on the frequency range most relevant to room modal behaviour while maintaining practical training efficiency at the reduced sampling rate of 480 Hz.

This sampling rate, though limiting for full-bandwidth applications, provided sufficient

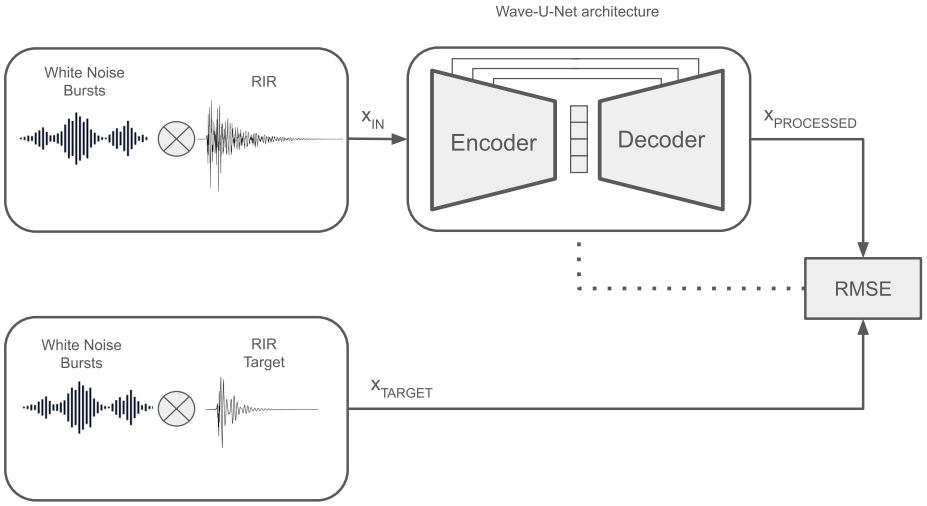


Fig. 1 - Overview of the training pipeline. The model processes an input signal (X_{IN}) and generates a predicted output ($X_{PROCESSED}$), which is compared to a target signal (X_{TARGET}) using time domain root mean square error (*RMSE*) as a performance metric.

[1] https://carlo-bolla.github.io/aes2025-ml-ai-audio/

resolution for the low-frequency modal control task while enabling reasonable computational requirements during training and final model size.
Such an architecture can be integrated as a component within a multi-rate filtering algorithm.

Fig. 1 illustrates the supervised learning framework employed for training the Wave-U-Net architecture. The input signals (X_{IN}) are synthesized by convolving modulated white noise burst (defined in Section 2.1) excitations with RIRs, while the corresponding target signals (X_{TARGET}) are generated through convolution of the same excitation signals with perceptually-optimized target RIRs. These target RIRs are derived from their reverberant counterparts by selectively reducing modal decay times to meet PMTs, ensuring that resonance suppression aligns with human auditory perception.
The framework leverages the commutative property of convolution operations, allowing the training process to focus on learning the mapping between resonant and perceptually-optimized acoustic responses while maintaining computational efficiency.

2.1 Excitation Signal

The excitation signals employed in the dataset consisted of amplitude-modulated white noise bursts, specifically designed to provide uniform spectral excitation across the entire target frequency range (20-240 Hz). White noise inherently provides full-spectrum stimulation with equal energy distribution across all frequencies.
To enhance dataset diversity and reduce the likelihood of the model learning fixed temporal patterns leading to potential overfitting [13], amplitude modulation was applied to the white noise bursts with randomized temporal parameters. These modulation parameters were systematically varied according to the following constraints:

$$\begin{cases} 0.3\ Seconds \leq L \leq 2.0\ Seconds \\ 0.01 \leq \dfrac{A}{L} \leq 0.20 \\ 0.10 \leq \dfrac{D}{L} \leq 0.80 \end{cases} \quad (1)$$

Where:

- L is the white noise burst length;
- A is the attack time;
- D is the decay time

The amplitude envelope transitions during attack and decay phases were shaped using a Blackman window function to eliminate abrupt discontinuities that could introduce spectral artifacts. This modulation approach generated a comprehensive range of excitation signals with varying temporal characteristics, potentially supporting improved model generalization to diverse resonant environments.
Multiple modulated bursts were concatenated to form extended sequences, further increasing dataset variability and enabling the model to process longer reverberant decay characteristics rather than isolated transient events.
To maintain consistency in the training pipeline, both input and target signals were standardized to 4-second durations. This temporal window was selected to ensure complete capture of modal decay characteristics within the target frequency range while balancing computational efficiency with the temporal resolution required for effective modal equalization analysis.
The 4-second duration provides sufficient temporal context for the model to analyse the full decay envelope of low-frequency room modes, which typically exhibit longer decay times compared to higher-frequency components.

2.2 Room Impulse Response

The synthetic RIRs were generated using modal decomposition formulated for cuboid rooms with rigid boundaries [11].
The modal accounts for modal pressure distributions, resonance frequencies, and damping effects based on the three-dimensional coordinates of both source and receiver positions. The source is modelled as an omnidirectional point source.
The room transfer function is formulated [12] as:

$$p_w(sr) = j\rho Q c^2 \sum_n \frac{p_n(s)p_n(r)}{K_n(\omega^2 - \omega_n^2 - 2j\delta_n\omega_n)} \quad (2)$$

Where:

- ρ is the air density $[kg/m^3]$;
- Q is the source strength $[m^3/s]$;
- c is the speed of sound $[m/s]$;
- n is the n^{th} order mode;
- $p_n(s)$ and $p_n(r)$ correspond to the modal pressure distributions at the receiver and source positions respectively;
- K_n is a mode-specific scaling factor;
- $\omega\ is$ the driving angular frequency of the source $[rad/s]$;
- ω_n defines the resonant angular frequency of the n^{th} mode $[rad/s]$;
- δ_n the modal damping coefficient,

The modal damping coefficient of a single resonant frequency, can also be expressed as a function of the modal decay as follows:

$$\delta_n = \frac{2.2\ \pi}{MD_{60,\mathrm{n}}} \cong \frac{6.91}{MD_{60,\mathrm{n}}} \tag{3}$$

Where $MD_{60,n}$ represents the modal decay time of the *n-th* resonant frequency ω_n.
It is computed as per [14] and depends on the room's absorption properties. The absorption coefficient α is assumed to be homogeneous across all room surfaces, meaning that all walls, floor, and ceiling have identical acoustic absorption properties.
The RIR target is derived with the same methodology except that $MD_{60,\mathrm{n}}$ is obtained from the refactored formulation of the PMTs [15]:

$$MD_{60,\mathrm{n}} = 0.15 + \frac{755}{f_n^2} \tag{4}$$

Where f_n is the resonant frequency of the n^{th} mode expressed in Hz.

The room configurations and placement rules were defined by the following parameters:

- Room dimensions: $\vec{L} = (L_x, L_y, L_z)$;
- Receiver position: $\vec{r} = (x_r, y_r, z_r)$;
- Source position: $\vec{s} = (x_s, y_s, z_s)$;
- Absorption coefficient α.

The room dimensions were randomly assigned based on uniform distributions, with the following constraint applied to prevent cubic or nearly cubic configurations:

$$\vec{L} = \begin{cases} 2.5\ m \le L_x < 6.1\ m \\ 2.5\ m \le L_y < 6.1\ m \\ 2.5\ m \le L_z < max(L_x, L_y) \end{cases} \tag{5}$$

Source constraints:

$$\vec{s} = \begin{cases} 0.15\ m \le x_s < L_x - 0.15\ m \\ 0.15\ m \le y_s < L_y - 0.15\ m \\ 0.30\ m \le z_s < 0.50\ m \end{cases} \tag{6}$$

Absorption coefficient:

$$0.05 < \alpha < 0.5 \tag{7}$$

2.3 Model Architecture

The proposed architecture (Fig. 2a) builds upon the Wave-U-Net framework [7], a fully convolutional neural network specifically designed for end-to-end time-domain audio processing.
The model employs an encoder-decoder structure with hierarchical downsampling and upsampling operations to efficiently extract multi-scale audio features while maintaining fine temporal resolution necessary for accurate modal frequency detection.
The encoder (Fig. 2b) consists of *M=5* convolutional blocks, each implementing a downsampling factor of 2 through stride-based 1D convolutions followed by PReLU activation functions. These stride-based convolutions progressively reduce the temporal resolution of the input signal, enabling the model to learn hierarchical representations of the audio.
The decoder structure (Fig. 2c), consists of N=5 symmetric blocks that systematically restore temporal resolution through learned upsampling operations. Each decoder block begins with feature map upsampling to match the corresponding encoder

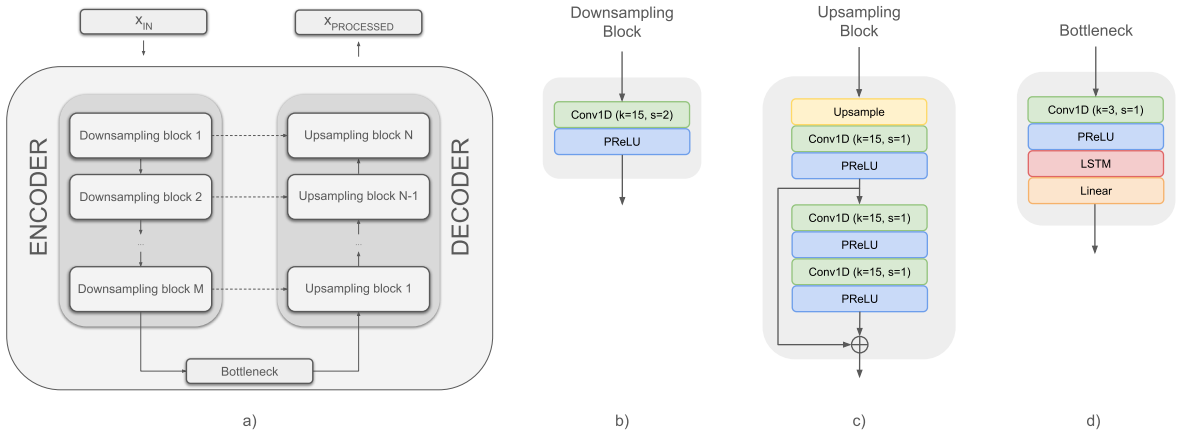


Fig. 2 - Model architecture schematics: a) Wave-U-Net, b) Encoder block, c) Decoder Block, d) Bottleneck.

resolution, followed by multiple 1D convolutional

layers with PReLU activations that progressively refine the reconstructed features. Skip connections merge decoder outputs with corresponding encoder features enhancing training stability.

At the bottleneck (Fig. 2d), a 1D convolutional layer with 144 channels further refines the learned representation, followed by a Long Short-Term Memory (LSTM*)* module with two layers, which enables the model to capture long-range dependencies in the signal to better accommodate low-frequency room mode equalization.

The addition of LSTM networks within the model addresses the need to capture long-range temporal dependencies inherent in processing the decay of low-frequency resonances. This integration has been explored in various studies, such as the work on low-latency live speech enhancement [16], where combining Wave-U-Net with LSTM modules improved the model's ability to handle temporal dynamics in speech.

Employing multiple convolutional layers per decoder block [17], enhances waveform reconstruction by progressively refining upsampled features, capturing temporal dependencies more effectively.

2.4 Training Process

The dataset was divided into training (80%), validation (10%), and test (10%) sets to ensure robust model generalization across the synthesized RIRs.

The data was fed into the model in batches of 64 samples, balancing computational efficiency with variability across different room conditions in each batch.

The model was optimized using the Root Mean Squared Error (RMSE) loss function computed in the time domain:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{PROCESSED,i} - x_{TARGET,i})^2} \qquad (8)$$

where N represents the total number of samples in each 4-second training sequence. This loss function was selected to minimize the temporal deviation between processed and target signals while preserving the overall energy characteristics of the modal suppression.

AdamW [18] optimization was employed due to its ability to adaptively adjust learning rates, enhancing stability during training.

The initial learning rate was set to 0.0001 and progressively decayed every 50 epochs to ensure smooth convergence. The model was trained for 250 epochs, with training performance monitored using the validation set.

To prevent excessive training durations and mitigate the risk of overfitting, an early stopping criterion was implemented. Training was automatically halted if the validation *RMSE* did not improve for 20 consecutive epochs.

3 Results

The performance of the modal control model was evaluated using dataset samples from the test subset, which were not included in training or validation, ensuring an unbiased assessment.

Fig. 3, Fig. 4 and Fig. 5 present results for three different room volumes excited with three different white noise bursts randomly generated according to Section 2.1.

The analysis focuses on the time-domain amplitude response and the frequency-domain magnitude response within the 20–240 Hz range. In all cases, the excitation signal was randomized to maintain consistency across test conditions (Section 2.1).

The time-domain amplitude response plots show the envelopes of the broadband signal, where the processed signal exhibits a reduction in the decay portion compared to the input signal while maintaining the attack of the white noise bursts.

The extent of decay attenuation varies across room volumes, with the processed signal following the target response closely in most cases.

Fluctuations in the processed signal are greater at smaller amplitudes (< -40 dBFS), which should be perceptually negligible.

A similar behaviour is observed in the frequency-domain magnitude response plots.

The resonant room modes in the input signal exhibit higher peaks, whereas the processed signal presents smoother peaks that more closely match the target response. In the high frequency regions, where higher modal order is observed, the processed spectra resemble a gain-adjusted version of the input signal, exhibiting similar features to the target magnitude response.

Additionally, on average, the magnitude values of the processed spectra are lower than the target, indicating a potential underestimation in gain compensation across frequencies. This suggests that the model adapts its spectral shaping to match the target while preserving the overall structure of the frequency response.

4 Discussion

The proposed model is successful in identifying and attenuating room modes, effectively reducing their

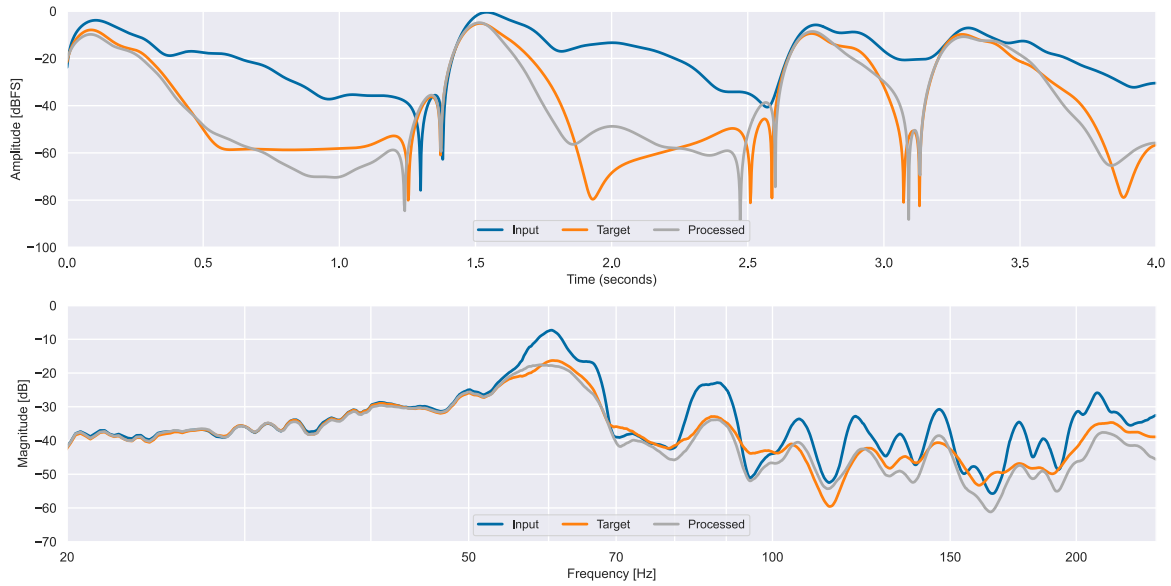


Fig. 4 – Room Volume 21 m^3 -Top: Time response (scaled in dBFS) of the Input (blue), Target (orange), and Processed (grey) envelopes.– Bottom: Frequency response of the Input (blue), Target (orange), and Processed (grey) signals.

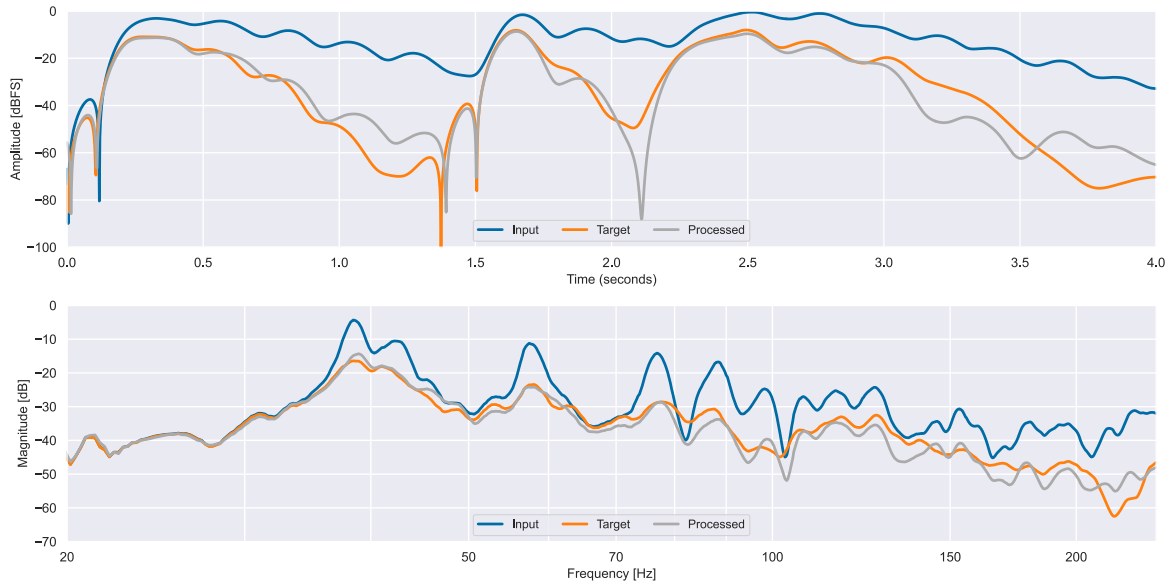


Fig. 3 – Room Volume 67 m^3 -Top: Time response (scaled in dBFS) of the Input (blue), Target (orange), and Processed (grey) envelopes.– Bottom: Frequency response of the Input (blue), Target (orange), and Processed (grey) signals.

decay to levels consistent with PMTs (orange curve in the time-domain plot). Traditionally, the detection and mitigation of room resonances rely on empirical methods, often requiring iterative adjustments. In contrast, this model provides a systematic and data-

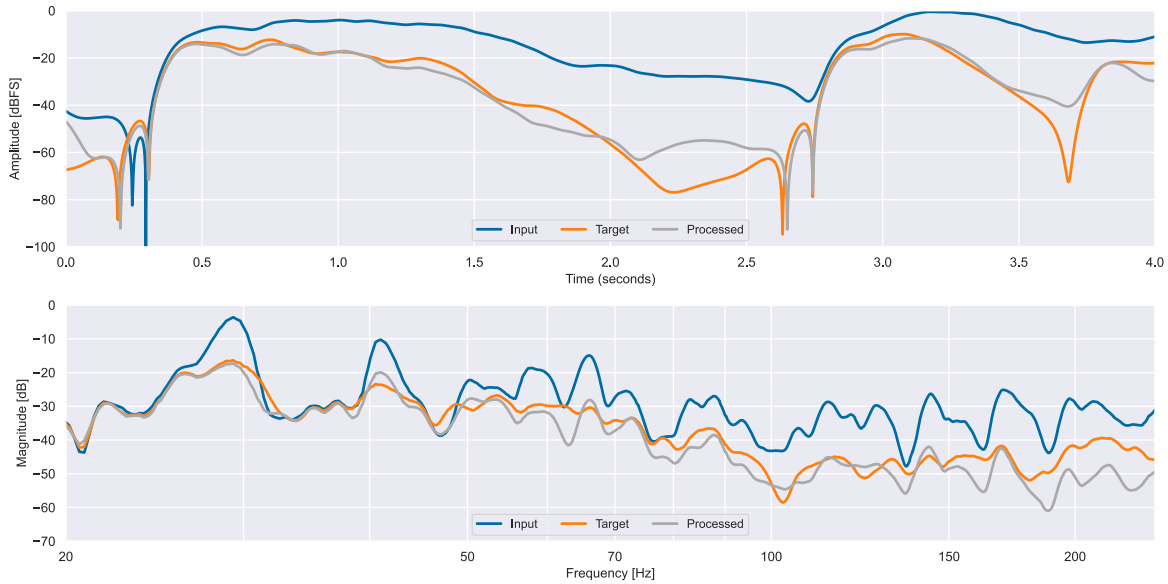


Fig. 5 – Room Volume 208 m^3 – Top: Time response (scaled in dBFS) of the Input (blue), Target (orange), and Processed (grey) envelopes.– Bottom: Frequency response of the Input (blue), Target (orange), and Processed (grey) signals.

driven approach to modal control, improving efficiency in room acoustics correction.

As shown in Fig. 3, Fig. 4 and Fig. 5, its performance remains consistent across the dataset, effectively adapting to different room volumes and reducing modal decay in accordance with the target response. The time domain correction is also noticeable in the frequency domain where spectral peaks attributable to room modes are reduced in magnitude and exhibit increased bandwidth.

Moreover, the envelopes of the processed white-noise signals, particularly the attack and decay segments, closely track those of the target, with only minor deviations.

Differences become apparent only at levels below −40 dBFS, where the processed output diverges slightly from the target. This behaviour arises from the use of RMSE as the loss criterion; the least-squares formulation minimises overall error and thus affords diminished weighting to very low-amplitude regions.

Consequently, while the overall spectral envelope is corrected, residual peaks persist at these frequencies. In contrast, for lower-order modal frequencies, the model effectively suppresses modal energy, leading to a smoother spectral response where the prominent peaks present in the input are no longer visible. This suggests that the model is particularly effective at low-frequency modal correction, while further refinements in the loss function could improve peak suppression in the higher-frequency range.

Although these peaks remain visible in the spectra, they may not be perceptually relevant, as room modes at these frequencies fall below the PMTs and the target spectral response.

Formal subjective listening tests may be required in future work to further investigate this aspect.

The generalization of the model to real-world signals, such as music and measured RIRs, remains an open research question. Since the model was trained exclusively on white noise bursts and synthesized RIRs, its performance on naturally occurring excitation signals is currently limited.

When applied to music or real RIRs, the model exhibits a notable reduction in accuracy, likely due to the spectral and temporal characteristics of the training data.

5 Conclusion

This study implemented a deep learning-based approach for low-frequency modal suppression in room acoustics using a modified Wave-U-Net architecture trained on synthetic RIRs.

Results show that the model effectively reduces modal decay times, aligning them with the target values as confirmed by the frequency-domain analysis, which indicates that low-order modal frequencies are successfully attenuated, resulting in a smoother spectral response compared to the input.

Higher-order modal frequencies exhibit residual spectral peaks, suggesting that the current model prioritizes overall spectral alignment but does not explicitly penalize localized deviations.
A refined loss function incorporating spectral smoothing constraints could improve spectral peak suppression while maintaining alignment with the target response.
However, their perceptual relevance remains uncertain and may require formal subjective listening tests for further assessment.
While the model performs as expected within the synthetic dataset, its generalization to real-world signals, such as music and measured RIRs, remains an open question.
Training was conducted using white noise bursts, ensuring uniform spectral excitation, but lacking the more idiosyncratic harmonic and temporal characteristics of musical signals.
When tested on non-synthetic signals, deviations in spectral alignment suggest that the model may not fully generalize to real-world excitation sources.
Future work will focus on retraining the model with musical excitation signals to improve its adaptability to real-world scenarios.
This transition introduces additional challenges due to the spectral sparsity, non-stationarity, and dynamic range variations of music, which differ significantly from white noise excitation. Data augmentation techniques and perceptually motivated loss functions will be investigated to enhance robustness.
Additionally, introducing causality into the model would improve its real-time applicability, making it more suitable for live audio processing and low-latency applications.

References

[1] A. Weisser and J. H. Rindel, "Evaluation of sound quality, boominess, and boxiness in small rooms.," *Journal of the Audio Engineering Society,* vol. 54, no. 6, pp. 495-511, 2006.

[2] F. A. Everest, Master Handbook of Acoustics, McGrawHill Education Tab, 2014.

[3] F. E. Toole, Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms, Routledge, 2017.

[4] T. Corbach, A. Von Dem Knesebeck, K. Dempwolf, M. Holters, P. Sorowka and U. Zölzer, "Automated Equalization For Room Resonance Suppression," in *Conference on Digital Audio Effects - DAFX*, Como, 2009.

[5] B. Owsinski , The Mixing Engineer's Handbook, Bobby Owsinski Media Group, 2022.

[6] M. Senior, Mixing Secrets for the Small Studio, Routledge, 2018.

[7] D. Stoller, S. Ewert and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation.," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, Paris, France, 2018.

[8] C. Macartney and T. Weyde, "Improved Speech Enhancement with the Wave-U-Net. 2018," in *https://arxiv.org/abs/1811.11307*, arxiv.org, 2018.

[9] J. Donley and P. Calamia, "DARE-Net: Speech Dereverberation And Room Impulse Response Estimation," *Stanford University,* 2022.

[10] B. M. Fazenda, M. Stephenson and A. Goldberg, “Perceptual thresholds for the effects of room modes as a function of modal decay,” *The Journal of the Acoustical Society of America,* vol. 137, no. 3, pp. 1088-1098, 2015.

[11] H. Kuttruff, Room Acoustics, CRC Press, 2019.

[12] B. Fazenda, M. R. Avis and W. J. Davies, "Perception of Modal Distribution Metrics in Critical Listening Spaces-Dependence on Room Aspect Ratios," *Audio Eng. Soc,* vol. 53, no. 12, p. 1128–1141, 2005.

[13] E. Kharitonov, R. Morgane, S. Gabriel, W. Lior, P.-E. Mazaré, M. Douze and E. Dupoux, "Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review," Shenzhen, China, 2020.

[14] J. H. Rindel, "Modal energy analysis of nearly rectangular rooms at low frequencies.," *Acta Acustica united with Acustica,* vol. 101, no. 6, pp. 1211-1221, 2015.

[15] J. Heddle, "Room acoustics for listening," in *2nd Australasian Acoustical Societies Conference*, 2016.

[16] B. Hanbin, P. Andreev, A. Saginbaev, N. Babaev, W.-J. Lee, H. Sung and H.-Y. Cho, "Speech Boosting: Low-Latency Live Speech Enhancement for TWS Earbuds," in *Interspeech*, Kos, Greece, 2004.

[17] M. T. Ho, J. Lee, B.-K. Lee, D. H. Yi and H.-G. Kang, "A Cross-Channel Attention-Based

Wave-U-Net for Multi-Channel Speech Enhancement," in *Interspeech*, Virtual conference, 2020.

[18] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, New Orleans, LA, USA, 2019.

[19] B. Katz, Mastering Audio: The Art and the Science, Routledge, 2014.

[20] S. Cecchi, A. Carini and S. Spors, "Room Response Equalization—A Review," Applied Sciences, 2018.

[21] J. Pons, S. Pascual, G. Cengarle and J. Serrà, "Upsampling Artifacts In Neural Audio Synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021.