Blind Acoustic Parameter Estimation and Spatial Audio Rendering Using Deep Learning for Augmented Reality

Petros Petrou
Student Number: 240910783
Supervisor: Dr. Aidan Hogg
Programme of Study: MSc FT Computer Science
Queen Mary University of London

***Abstract*—This thesis presents a complete end-to-end system for blind acoustic parameter estimation and spatial audio rendering in augmented reality (AR) environments. The AR pipeline enables real-time, scene-aware audio synthesis without requiring manual room measurements or specialized equipment. A deep learning model based on a Convolutional Recurrent Neural Network (CRNN) is trained on synthetically generated reverberant speech to predict key acoustic parameters directly from mono audio. These parameters are then used to drive a parametric rendering module based on Feedback Delay Networks (FDNs), producing binaural audio consistent with the estimated room acoustics. Extensive evaluation, including objective metrics (MAE, Pearson's r, PESQ, STOI) and subjective listening tests, demonstrates that the system reliably estimates room characteristics and synthesizes perceptually realistic, intelligible spatial audio. The results highlight the feasibility of blind acoustic modeling as a scalable front-end for immersive AR audio applications.**

I. Introduction

As augmented reality (AR) technologies continue to evolve, achieving perceptually realistic audio rendering is becoming as critical as the visual experience. For an immersive AR experience, spatial audio must reflect the acoustics of the physical environment, including reverberation, reflections, and spatial cues. A core component of spatial sound rendering is the Room Impulse Response (RIR), which characterizes how sound propagates in an environment from a source to a listener. However, measuring RIRs manually or using calibrated hardware is time-consuming, impractical for mobile applications, and incompatible with the real-time demands of AR.

To address this, the present work investigates the problem of blind estimation of acoustic parameters that describe the characteristics of a room's impulse response—such as reverberation time (RT60), direct-to-reverberant ratio (DRR), and clarity (C50)—directly from speech or audio signals. This approach does not attempt to recover the full RIR waveform, but focuses instead on estimating perceptually and physically meaningful parameters derived from it. If accurate, these blind estimates can drive spatial audio rendering in AR with minimal user intervention, enabling scalable and real-time deployment.

Recent research has shown that deep learning models, particularly convolutional and recurrent architectures, can effectively learn to estimate such acoustic parameters from mono audio recordings. Furthermore, advancements in transformer-based architectures have introduced powerful temporal modeling capabilities for this task. However, most of these systems remain disjointed from actual AR pipelines, and there is a gap in translating estimated acoustic parameters into perceptually compelling audio experiences.

In this work, that gap is addressed through the development of a deep learning model capable of blind estimation of key acoustic parameters and the demonstration of their integration within an immersive audio rendering context. The broader goal is to support scene-aware AR experiences by enabling perceptually realistic soundscapes through estimated room acoustics—without requiring expensive measurements or specialized devices.

II. Related Work

This section explores the state-of-the-art in room acoustics, spatial audio, and the use of deep learning in blind acoustic parameters estimation. The discussion is framed around the core goal of this project: enabling immersive and perceptually realistic spatial audio in AR.

The chapter begins by introducing foundational acoustic parameters, such as RT60, DRR and C50 before surveying the evolution of blind estimation methods—from statistical approaches to deep neural networks—as part of the background research. It then explores how estimated parameters are used in immersive audio rendering, particularly through techniques such as Scene-Aware Audio Rendering and AV-NeRF. Finally, it presents theoretical evaluation frameworks and identifies current research gaps. The structure follows a thematic progression with chronological elements embedded within each theme.

A. Fundamentals of Room Acoustics and Spatial Audio

Room Impulse Responses (RIRs) are temporal filters that capture the unique acoustic signature of a space. An RIR represents the response of a room to a short pulse, incorporating reflections, diffractions, and absorptions that occur as sound travels from a source to a receiver.

Key acoustic parameters derived from RIRs include RT60, DRR, and the clarity index C50. RT60 represents the time it takes for sound energy to decay by 60 dB, indicating the strength of reverberation in a space. The DRR quantifies the ratio between the energy of the direct sound and that of the reverberant field, offering insight into the spatial characteristics of the environment. C50 measures how much energy arrives within the first 50 milliseconds, which is closely related to speech intelligibility.

In immersive applications like augmented reality (AR), perceptual constructs such as clarity, externalization, and envelopment are essential. Binaural rendering using Head-Related Transfer Functions (HRTFs) captures spatial information that helps the human auditory system perceive the direction and spatial context of a sound.

B. Blind Acoustic Parameter Estimation

1) Early Approaches: Initial methods relied on handcrafted audio features and classical statistical models. However, these techniques were limited by generalizability and noise robustness.

2) CRNN-Based Models: Convolutional Recurrent Neural Networks (CRNNs) introduced temporal modeling capabilities. Callens & Cernak (2020) and López et al. (2021) proposed universal estimators that predict multiple acoustic parameters (e.g., RT60, DRR, C50) jointly from reverberated speech, achieving robustness to noise and different signal types.

3) Transformer-Based Models: More recent transformer models, such as those introduced by Wang et al. (2024*a*), utilize self-attention to capture long-range dependencies in the time-frequency space. These architectures improve performance, especially in complex environments with dynamic acoustic properties.

4) Phase-Based Features: Ick et al. (2023) explored the use of phase-sensitive spectrogram representations, demonstrating improvements in estimating reverberation time and clarity by preserving phase continuity.

5) Multichannel Models: Srivastava et al. (2021) extended blind estimation to multichannel inputs, enhancing the accuracy of spatial parameter predictions by capturing more room geometry cues.

C. Rendering Techniques for Immersive Audio

1) Scene-Aware Rendering: Tang et al. (2020) introduced techniques that use semantic scene understanding and geometric information to enhance sound spatialisation in AR environments.

2) Audio-Visual Neural Radiance Fields (AV-NeRF): Liang et al. (2023) proposed AV-NeRF, a novel method that integrates NeRF-based rendering with audio generation. Their model creates view-consistent binaural audio by encoding the spatial geometry and material properties of the environment into neural fields.

3) Reverberation Synthesis with FDNs and SDNs: Fundamental reverberation structures such as Feedback Delay Networks (FDNs) and Scattering Delay Networks (SDNs) are also employed in parametric spatial audio rendering. These are covered extensively in Katz & Majdak (2022) foundational text.

D. Evaluation Approaches

Several reviewed studies, including López et al. (2021), Wang et al. (2024*b*), and Katz & Majdak (2022), emphasize the need to evaluate both the accuracy of estimated parameters and how realistic the resulting audio sounds.

Objective methods such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Pearson's correlation coefficient (ρ) are used to compare predicted values to ground truth data.

Perceptual evaluation is also discussed in the literature. Some studies apply listener tests, such as MUSHRA (a method defined by ITU-R), to compare different versions of audio and rate their quality. Others focus on listener impressions of externalization, spatial clarity, and overall realism.

Despite these efforts, a consistent issue noted across papers is the lack of a standard framework that connects estimation errors with actual perceptual quality.

E. Research Gaps Identified

Despite recent advancements, the literature reveals several notable gaps. First, a modular, end-to-end pipeline that spans from speech input to immersive audio rendering remains underdeveloped, limiting the integration of estimation and playback systems. Second, evaluation methods frequently lack perceptual grounding; numerical accuracy alone does not necessarily reflect how realistic or immersive the resulting audio sounds to human listeners.

III. Methodology

A. Introduction to Methodology - Proposed System

This section details the methodology followed in designing, implementing, and evaluating a pipeline system as shown in Figure 1 for blind acoustic parameters estimation and immersive spatial audio rendering in the context of AR — without requiring manual measurements or specialized equipment.

To achieve the goals of the project, the work is divided into two major phases:

Phase I - Blind Acoustic Parameters Estimation: In this phase (Figure 2), a deep learning model based on a CRNN is developed to predict key acoustic parameters—RT60, DRR, and C50—directly from reverberant speech signals. The model is trained using a synthetically generated dataset, where each sample is produced by convolving clean speech with simulated room impulse responses using the PyRoomAcoustics toolkit (Scheibler et al. (2018)).

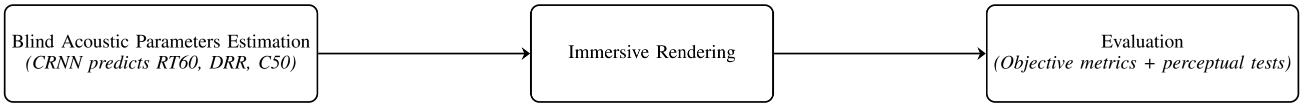

Blind Acoustic Parameters Estimation
(CRNN predicts RT60, DRR, C50)
Immersive Rendering
Evaluation
(Objective metrics + perceptual tests)

Fig. 1: High-level architecture of the proposed pipeline: blind estimation of room acoustics, spatial audio rendering, and evaluation of realism.

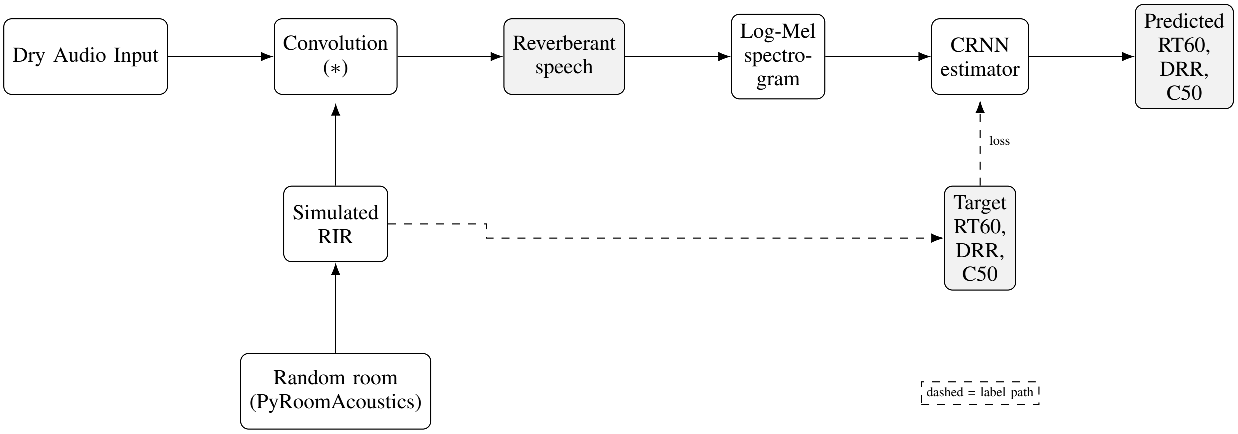


Fig. 2: Estimation phase: synthetic reverberant speech is generated using PyRoomAcoustics and fed into a CRNN to predict RT60, DRR, and C50. Ground-truth labels from the simulated RIR supervise the training.

Phase II - Spatial Audio Rendering: The second phase (Figure 3) uses the predicted parameters to synthesize immersive spatial audio. This involves generating spatialized reverberation effects consistent with the estimated room characteristics, by using an FDN. This stage is critical for integrating the system into AR pipelines, enabling dynamic and perceptually plausible soundscapes without physical room measurements.

B. Phase I - Blind Acoustic Parameters Estimation

1) Synthetic Dataset Generation: To train the deep learning model for blind acoustic parameter estimation, a large-scale synthetic dataset was constructed. The purpose of that is to allow precise control over room conditions and parameter labels. The dataset is consisted with reverberant speech samples paired with analytically derived acoustic parameters.

Clean Speech Source: The clean speech samples are sourced from the publicly available LibriSpeech corpus, specifically the *train-clean-100* subset as produced by Panayotov et al. (2015). A random selection of 50 clips was used to ensure coverage of speaker variation. Each clip was resampled to 16 kHz, converted to mono to simplify the task, and trimmed to a fixed duration of 5 seconds to standardize input length and limit computational cost during convolution.

Room Simulation: Each speech sample was convolved with a synthetically generated RIR using the PyRoomAcoustics simulation toolkit. This approach was followed in order to gain control over room geometry, absorption characteristics, and source-receiver placement. For each of the 10,000 training examples:

- Room dimensions were randomly sampled from:
  - Width/Length: $x, y \sim \mathcal{U}(3,\ 10)$ m
  - Height: $z \sim \mathcal{U}(2.5,\ 4)$ m
- A target reverberation time: $T_{60}^{*} \sim \mathcal{U}(0.2,\ 1.5)$ s
- The inverse Sabine formula was used to derive the corresponding wall absorption coefficients. This inversion ensures that the simulated room achieves the desired reverberation time, making the dataset controllable and well-balanced across acoustic conditions.
- A single microphone and single speech source were randomly placed inside the room, with a 0.5 m safety margin from walls.

Feature Extraction: For each reverberant signal, a log-Mel spectrogram was extracted using the *librosa* library (McFee et al. (2015)). The parameters included to the spectrogram were the 64 Mel bands, a window length of 25 ms and a hop size of 10 ms. In addition to this, the time axis of the spectrograms were padded or truncated to 160 frames to ensure fixed-size representation, suitable for batch processing. Moreover, each spectrogram was then standardised per sample to zero mean and unit variance.

Acoustic Parameter Targets: The target of the deep learning model is to calculate the following three parameters:

- T_{60}: Reverberation Time
- DRR: Direct-to-Reverberant Ratio
- C_{50}: Speech Clarity Index (50 ms window)

Each of these parameters was whitened using predefined statistics (mean and standard deviation) to stabilize training. The result was a three-dimensional vector of normalised targets associated with each input sample.

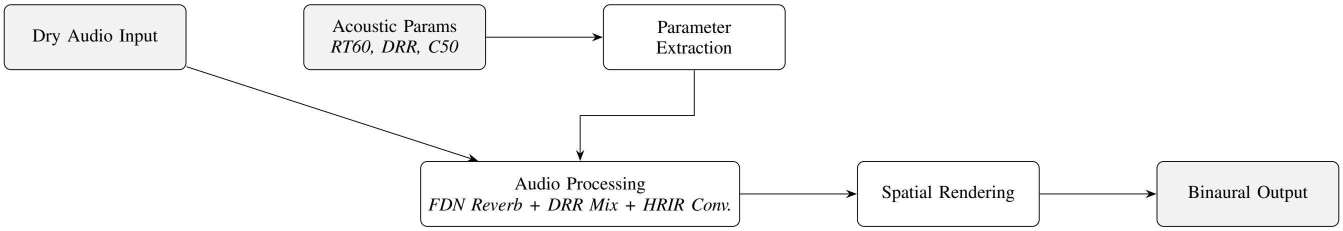


Fig. 3: Implementation pipeline: dry audio and predicted acoustic parameters are used to generate spatialized binaural output using an FDN-based reverberation model and HRIR convolution.

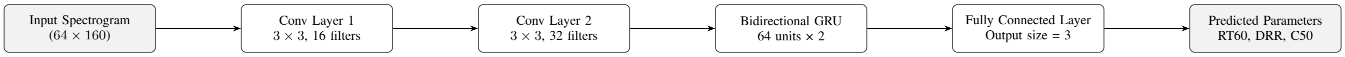


Fig. 4: CRNN architecture used for blind room acoustics estimation. Log-Mel spectrograms are passed through convolutional layers to extract time-frequency features, followed by a bidirectional GRU for temporal modeling, and a fully connected layer that predicts RT60, DRR, and C50.

2) Model Architecture: To estimate acoustic room parameters from reverberant speech, the model must learn both frequency-based patterns and how these patterns evolve over time. For this purpose, a CRNN architecture was employed. This model combines two types of layers: convolutional layers for local feature extraction in time-frequency space, and recurrent layers for modeling the temporal progression of reverberation effects. The proposed architecture is shown in Figure 4 . This design aligns with prior work in the field, such as the estimator proposed by López et al. (2021).

Convolutional Front-End: The model receives as input a log-Mel spectrogram of size *64 × 160*, where 64 represents Mel frequency bands and 160 is the number of time frames. This input is passed through two convolutional layers. The first convolutional layer uses 16 filters of size *3 × 3*, followed by a ReLU activation and a *2 × 2* max pooling operation. These filters act as pattern detectors over small time-frequency patches of the spectrogram, while the max pooling operation reduces dimensionality and retains the most prominent features. The second layer increases the number of filters to 32 and follows the same structure, allowing the network to capture more abstract and high-level features. These convolutional layers transform the spectrogram into a set of compressed feature maps that highlight acoustic cues relevant to reverberation.

Temporal Modeling with Bidirectional GRU: After the convolutional layers, the output is reshaped into a sequence of time steps, where each step contains a set of extracted features. This sequence is then fed into a Bidirectional Gated Recurrent Unit (GRU), which reads the data in both forward and backward directions. By doing so, the model can understand how sounds evolve over time and how they relate to what came before and after. This is especially useful for predicting parameters like RT60 and C50, which rely on the timing and decay of sound in the signal. The GRU uses 64 hidden units in each direction, producing a combined output size of 128 features.

Output Layer: Finally, a fully connected layer maps the GRU output to a three-dimensional vector corresponding to the predicted acoustic parameters: RT60, DRR, C50. No activation function is used at this stage since the targets are continuous values.

3) Training Configuration:

Dataset Preparation & Splitting: The dataset was randomly split into 80% training and 20% validation subsets. To maintain consistent evaluation, the splits were fixed across training runs.

Loss Function: The model was trained to minimize the Mean Squared Error (MSE) loss between predicted and target values:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{3}\sum_{i=1}^{3}(\hat{y}_i - y_i)^2 \tag{1}$$

This loss was computed on the whitened targets, which standardize the parameter scales and stabilize gradient updates during training. As a result, each parameter contributes equally to the overall loss regardless of its physical scale.

Optimizer and Hyperparameters: The model was optimized using the *Adam* optimizer, which adapts learning rates based on estimated first- and second-order moments of the gradients. The following settings were used:

- Learning rate: 1×10^{-3}
- Batch size: 32
- Number of epochs: 30
- No learning rate scheduler was applied, as performance was stable with a constant rate.

4) Evaluation of the Estimator: Four regression-based metrics were used to quantify performance across the three predicted acoustic parameters.

The *Mean Absolute Error (MAE)* measures the average absolute deviation between predicted and target values in the physical domain:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{2}$$

The *Root Mean Squared Error (RMSE)* is similar to MAE but places greater emphasis on larger errors due to squaring.

The *Pearson's Correlation Coefficient (r)* assesses the strength of the linear relationship between predicted and true values. A higher correlation indicates that the model successfully captures consistent trends, even if there is a scale offset.

Finally, the *Coefficient of Determination (R^2)* indicates the proportion of variance in the target variable that is explained by the model. An R^2 value close to 1 implies strong predictive accuracy.

5) Real-World Dataset Integration: In addition to training and evaluating the estimator on synthetic data, the system was also applied to recordings from the BUT Speech@FIT Reverb Database Szöke et al. (2019). This dataset includes speech from the LibriSpeech corpus Panayotov et al. (2015) convolved with real RIRs measured in a variety of real environments. Among these, the Q301 office subset was selected for inference due to its balanced acoustics and structured data organization. The room has a volume of 192 m³ (10.7 m × 6.9 m × 2.6 m) and includes reverberant speech recordings from 31 positions and 3 microphone placements.

The predictions were served as input to the spatial rendering pipeline described in Phase II. This step forms a key bridge between blind acoustic estimation and practical audio rendering using real-world data.

C. Phase II - Spatial Audio Rendering

1) Overview: After estimating the acoustic parameters in Phase I, the goal of this stage is to use those values to generate spatial audio that matches the characteristics of the predicted room. This section presents the system used to render binaural audio based on the estimated RT60, DRR, and C50 values. The rendering process combines reverberation, spatial filtering, and parameter-driven control to simulate a realistic and immersive audio scene, as shown in Figure 3. To achieve this, the system employs a parametric reverberation model based on an FDN. The resulting reverberant signal is then spatialized using HRIRs, producing binaural audio consistent with natural human spatial perception.

2) Motivation for Using FDN: Traditional reverberation methods, such as simple exponential decay functions, have been widely used due to their ease of implementation and control. However, as demonstrated in early work by Schroeder (Schroeder (1962)), such models do not fully capture the complexity of real acoustic environments. In particular, they often fail to produce natural-sounding reverberation tails, which can feel artificial or distracting—especially in immersive applications like virtual or augmented reality, where perceptual realism is critical.

To address these limitations, an FDN was chosen as the core of the rendering system. FDNs simulate multiple interconnected delay lines with feedback, allowing them to generate a more diffuse and natural-sounding reverberation tail. While the implementation remains relatively lightweight, the feedback structure introduces time-varying energy buildup and complexity that more closely approximates sound propagation in reflective spaces.

3) System Design and Implementation: The rendering system takes the predicted acoustic parameters—RT60, DRR, and C50—and uses them to produce a spatialized audio signal that reflects the estimated reverberant environment. This process involves three main stages: reverberation synthesis using an FDN, dry-to-reverberant mixing based on DRR, and spatialization using HRIRs, as shown in Figure 3.

The pipeline starts with a monophonic dry speech waveform and a parameter triplet obtained from the CRNN estimator in Phase I. Among the predicted values, RT60 and DRR are actively used in the rendering process. C50, although estimated, is treated as an evaluation metric and not directly involved in signal transformation.

Reverberation is applied using a custom-built FDN reverb module. This consists of four fixed prime-length delay lines (149, 211, 263, and 293 samples), each fed back with a gain coefficient derived from the predicted RT60. These gain values are calculated to simulate an energy decay that matches the desired reverberation time. The feedback structure allows energy to circulate and diffuse across the delay lines, producing a denser and more natural reverberant tail compared to a simple exponential decay.

Once the reverberant signal is generated, it is scaled and added to the dry signal using a gain ratio derived from the predicted DRR (converted from dB to linear scale). This simulates the energy balance between the direct and reflected sound components in the room. The result is a reverberant but still intelligible waveform that reflects both temporal decay and energy proportion.

To convert the processed mono signal into a spatial binaural output, the system uses HRIRs from the CIPIC HRTF database Algazi et al. (2001), selecting the filter pair corresponding to azimuth 0° and elevation 0° for a straightforward front-facing spatialization. The dry + reverb signal is convolved separately with the left and right HRIR filters to simulate directional cues based on human head and ear filtering effects. Finally, the two-channel waveform is normalized and saved as a stereo file for evaluation and playback.

4) Output Generation: The final output of the rendering system consists of stereo binaural audio files generated using predicted acoustic parameters from real-world speech samples. A total of 50 utterances were processed using the FDN-based pipeline, producing perceptually enhanced spatial audio consistent with the estimated room characteristics.

5) Evaluation Plan: To assess the quality and plausibility of the rendered outputs, a combination of objective and informal subjective methods was used. The goal is to evaluate whether the generated binaural audio is consistent with the estimated acoustic parameters and maintains clarity and realism.

For objective evaluation, two established perceptual metrics are used: PESQ (Perceptual Evaluation of Speech Quality), which estimates the impact of reverberation on speech quality, and STOI (Short-Time Objective Intelligibility), which measures how intelligible the processed speech remains in comparison to the clean input. Both metrics are computed using the dry speech signal as reference and are applied to the outputs of the FDN-based system.

In addition, acoustic consistency is assessed by estimating RT60 and C50 from the rendered signals and comparing them to the predicted values. These provide a check on whether the decay time and clarity characteristics are preserved after rendering.

The evaluation also includes visual inspection of waveform envelopes and energy decay curves to support qualitative analysis of the reverberant tail structure. In addition, a small number of representative examples were evaluated through subjective listening and informal commentary, providing further insight into the perceived spatial impression and audio quality.

IV. Evaluation Results

A. *Evaluation of Phase I*

The evaluation of the estimator was carried out on the synthetic dataset using a fixed training and validation split. The performance was assessed using standard regression metrics, along with visualizations to better interpret the model's behavior.

Figure 5 shows the training and validation loss curves across 30 epochs. The model converges rapidly within the first few epochs, and the validation loss closely tracks the training loss throughout. This smooth and consistent behavior is indicative of stable training dynamics and generalization without overfitting.

Quantitative performance is summarized in TABLE I, where the MAE, RMSE, R^2 score, and Pearson correlation coefficient are reported for each of the three target parameters. RT60 exhibited strong correlation with ground-truth values (r = 0.968), although the absolute prediction error was higher than for the other targets. In contrast, DRR and C50 were predicted with notably lower errors (MAE = 0.59 dB and 0.15 dB, respectively) and extremely high correlation values, particularly in the case of DRR, which achieved a Pearson r of 0.996. These results demonstrate the model's ability to learn both energetic and temporal characteristics of reverberation effectively.

To further support the quantitative findings, Figure 6 illustrates scatter plots comparing the predicted and ground-truth values for all three parameters. The plots confirm the high alignment between estimates and targets, with the majority of points clustering tightly around the diagonal reference line. While RT60 shows slightly more variance in predictions, the DRR and C50 plots reflect high accuracy and strong linearity across the value range.

In summary, the estimator demonstrates reliable performance across all three target parameters. The combination of low prediction error and high correlation confirms that the trained CRNN was able to learn a robust mapping. From time-frequency input features to key acoustic characteristics, validating the use of the synthetic dataset and the model architecture for this task.

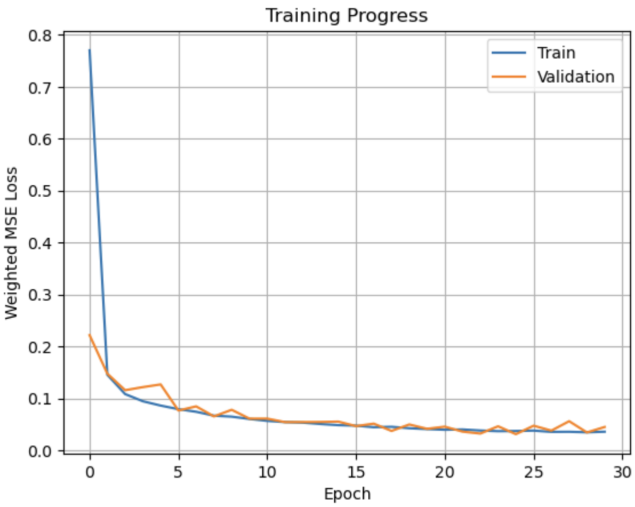


Fig. 5: Model Training Progress

TABLE I: Evaluation metrics on the synthetic validation set for each predicted parameter.

Metric	MAE ↓	RMSE ↓	R^2 ↑	Pearson r ↑
RT60	1.5286 s	1.9410 s	0.9087	0.9680
DRR	0.5938 dB	0.7415 dB	0.9888	0.9964
C50	0.1514 dB	0.2180 dB	0.8995	0.9640

B. *Evaluation of Phase II*

To evaluate the perceptual quality and intelligibility of the rendered audio, two standard objective metrics were used: PESQ and STOI. These metrics assess how closely the spatialized audio resembles the original dry input, in terms of perceived quality and speech clarity respectively.

The evaluation was conducted on 50 FDN-rendered audio samples, using the corresponding dry (anechoic) recordings as references. As required by both metrics, all signals were downmixed to mono. Each rendered signal was paired with its corresponding dry waveform, and both PESQ and STOI were computed using their standard Python implementations.

The distribution of PESQ scores is shown in Figure 7a. Most samples achieved scores between 2.5 and 3.5, with a clear peak around 3.0. This range is consistent with moderate-to-high perceived quality in blind spatialization tasks, where some spectral and temporal distortion is expected due to reverberation synthesis. The mean PESQ score was 2.99, with a standard deviation of 0.20, indicating relatively consistent performance across the dataset.

Similarly, the distribution of STOI scores is presented in Figure 7b. The scores were tightly clustered near 1.0, with most samples exceeding 0.95. The average STOI score was 0.96, suggesting that the rendered audio preserved most of the intelligibility of the original dry signals, even after convolution with reverberant and spatial effects.

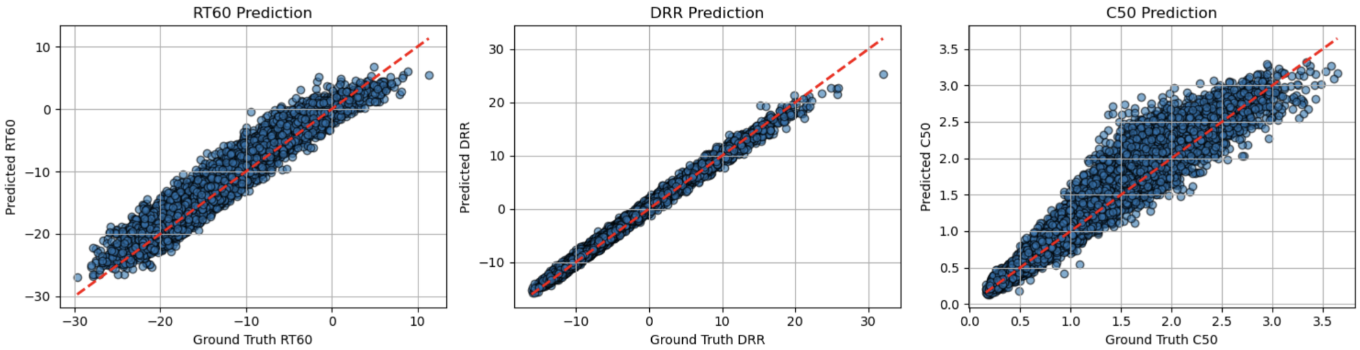
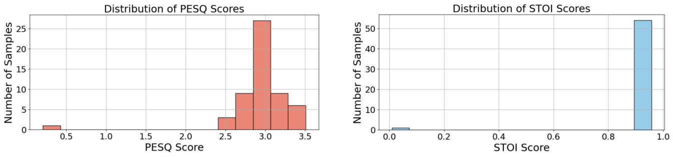


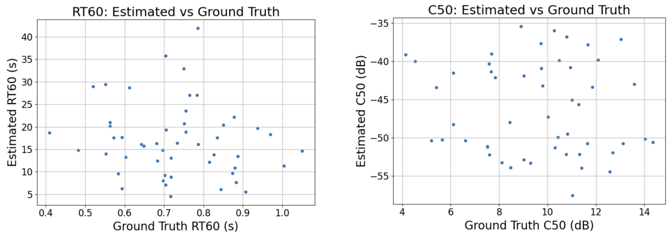
Fig. 6: Predicting vs Ground Truth Scatter Plots



(a) PESQ Histogram

(b) STOI Histogram

Fig. 7: Objective evaluation histograms for perceptual quality (PESQ) and intelligibility (STOI).



(a) RT60: Estimated vs Ground Truth

(b) C50: Estimated vs Ground Truth

Fig. 8: Predicted vs. ground truth values for RT60 (a) and C50 (b) on real-world speech data.

These results confirm that the FDN rendering pipeline produced outputs that are perceptually close to the reference signals and intelligible to a high degree. The high STOI scores in particular indicate that speech clarity was maintained, which is especially important in applications involving augmented or virtual reality environments.

Acoustic Consistency: While perceptual metrics such as PESQ and STOI are valuable for assessing subjective audio quality, they do not guarantee that the physical characteristics of the rendered signal remain consistent with the predicted acoustic parameters. To address this, a separate analysis was performed that directly examines whether the reverberation in the FDN-rendered outputs aligns with the estimated RT60 and C50 values provided to the system.

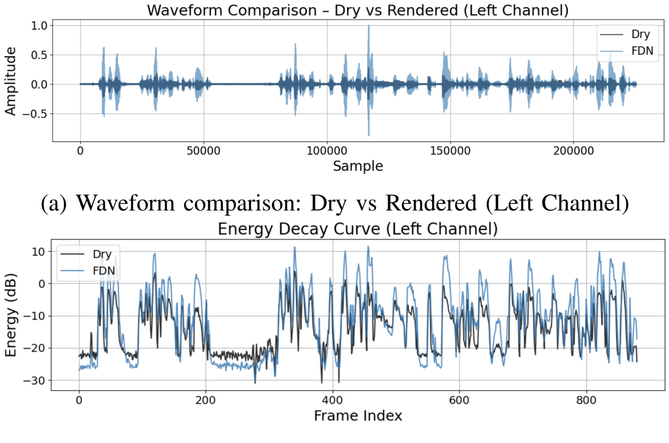
RT60 was computed from the rendered audio using an energy decay curve (EDC) method, which estimates the reverberation time based on the logarithmic drop in signal energy over time. Similarly, C50 was estimated by measuring the early-to-late energy ratio in the time domain, using a 50 ms integration window as standard.

Figure 8a compares the estimated RT60 values computed from the rendered signals with the original predicted RT60 values that were used as input to the rendering stage. The estimated reverberation times fall within a plausible range and demonstrate a loose but recognizable alignment with the input values. A similar trend is seen in Figure 8b, where the estimated C50 values (in dB) are plotted against their corresponding predicted targets. Despite a relatively wider spread, the estimates still cluster within a meaningful range, reflecting the reverberation clarity introduced by the FDN system.

These comparisons are not meant to be interpreted as strict regression metrics but rather as plausibility checks. The key takeaway is that the rendered reverberation exhibits broadly consistent decay and energy behavior relative to the expected room characteristics, affirming the effectiveness of the FDN rendering pipeline as a physically informed synthesizer of acoustic properties.

Subjective Analysis: To complement the objective analysis, a subjective evaluation was conducted to qualitatively assess the realism and perceptual plausibility of the rendered audio. This included visual inspection of waveform and energy decay behavior, as well as informal listening and rating of selected samples. Figure 9a presents a waveform comparison between the dry input signal and its corresponding FDN- and exponentially-rendered versions.

In addition to visual inspection, a small-scale subjective listening experiment was conducted, evaluating a subset of 25 binaural outputs generated using the FDN-based rendering system. For each sample, two key perceptual dimensions were rated: the perceived quality of the rendered audio and the realism of the reverberation effect. Ratings were assigned on a 5-point Likert scale, where one (1) indicated the best possible outcome and five (5) the worst. The results, aggregated from the annotated evaluation forms, showed that 64% of samples were rated as either “good” (2) or “fair” (3) in terms of perceived audio quality, with only 12% receiving scores of “poor” (4) or “very poor” (5). Similarly, the realism of the



(a) Waveform comparison: Dry vs Rendered (Left Channel)

(b) Energy decay curve (Left Channel)

Fig. 9: Temporal and energetic comparison between dry audio and rendered outputs. (a) Waveform structure across methods. (b) Energy decay behavior.

reverberation was judged positively in the majority of cases, with 60% of examples rated 2 or 3, reflecting a strong correspondence between the rendered reverberation and the acoustic expectation set by the dry reference.

These subjective assessments align well with the earlier objective findings. The FDN-generated audio not only maintains high intelligibility and perceptual quality metrics (PESQ and STOI), but also exhibits visual and auditory cues that reinforce its acoustic plausibility. Together, these results demonstrate that the pipeline successfully transforms estimated acoustic parameters into immersive spatial audio experiences that are not only intelligible but also realistic to the listener.

V. Conclusion

This thesis presented a complete end-to-end system for blind room acoustics estimation and immersive audio rendering, designed to operate in augmented reality contexts without requiring manual room measurements or specialized calibration. The proposed pipeline was divided into two phases: a deep learning model for blind estimation of key acoustic parameters (RT60, DRR, and C50), and a rendering module that synthesizes spatialized binaural audio using a parametric reverberation structure based on FDNs.

The CRNN-based estimator achieved high prediction accuracy on a synthetically generated dataset, demonstrating strong correlation with ground-truth values and low absolute error across all three target parameters. These results confirmed the model's ability to learn both the energetic and temporal aspects of room reverberation from log-Mel spectrogram inputs. Evaluation plots and loss curves further showed stable training behavior and generalization to unseen synthetic data.

Building on this, the second phase translated the estimated parameters into spatialized reverberant audio. The FDN implementation provided physically plausible and perceptually convincing reverberation behavior. Objective evaluations using PESQ and STOI indicated that the FDN-rendered signals preserved speech quality and intelligibility to a high degree. Additional analysis confirmed that the temporal decay patterns and energy distributions of the rendered signals were consistent with the expected room characteristics.

Subjective assessments further supported these findings. Informal listening tests, along with waveform and energy decay visualizations, showed that the FDN-based rendering produced immersive audio experiences that were both natural and intelligible. While some variation existed in the perceptual feedback, the overall impressions aligned strongly with the system's design goals.

Together, these results demonstrate that blind estimation of acoustic parameters can serve as a reliable front-end for generating spatial audio that is both physically informed and perceptually compelling. The system offers a scalable and deployable solution for realistic sound synthesis in AR environments, bridging a key gap between acoustic modeling and interactive audio rendering.

VI. Future Work

While this thesis presents a complete pipeline for blind acoustic parameter estimation and spatial audio rendering, several directions remain open for future exploration.

Firstly, the current model was trained on synthetic data with controlled acoustic properties. Incorporating larger and more diverse real-world datasets—especially those with known ground-truth room parameters—could improve robustness and support better generalization across different environments. Additionally, augmenting the input with phase-based features or multichannel recordings may improve the model's sensitivity to spatial cues and reverberation complexity.

On the rendering side, the FDN implementation could be extended to include frequency-dependent decay control, which would allow for finer tuning of the reverberation spectrum. Integration with dynamic scene understanding or head-tracked spatialization could also improve realism in real-time AR scenarios. Furthermore, the estimated C50 parameter, while used here primarily for evaluation, could be directly incorporated into the rendering algorithm to better control clarity and early reflections.

Lastly, while this work focused on informal subjective evaluation, a formal user study would help quantify perceptual differences between rendering methods and validate listener impressions at scale. This would be particularly relevant for applications in AR audio, where realism and immersion are critical.

References

Algazi, V. R., Duda, R. O., Thompson, D. M. & Avendano, C. (2001), The cipic hrtf database, *in* 'Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)', IEEE, pp. 99–102.

Callens, P. & Cernak, M. (2020), 'Joint blind room acoustic characterization from speech and music signals using convolutional recurrent neural networks', *arXiv preprint arXiv:2010.11167* .

Ick, C., Mehrabi, A. & Jin, W. (2023), Blind acoustic room parameter estimation using phase features, *in* 'ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 1–5.

Katz, B. F. & Majdak, P., eds (2022), *Advances in Fundamental and Applied Research on Spatial Audio*, IntechOpen.

Liang, S., Huang, C., Tian, Y., Kumar, A. & Xu, C. (2023), 'Av-nerf: Learning neural fields for real-world audio-visual scene synthesis', *Advances in Neural Information Processing Systems* **36**, 37472–37490.

López, P. S., Callens, P. & Cernak, M. (2021), A universal deep room acoustics estimator, *in* '2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)', IEEE, pp. 356–360.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E. & Nieto, O. (2015), librosa: Audio and music signal analysis in python, *in* 'Proceedings of the 14th python in science conference', Vol. 8.

Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. (2015), Librispeech: an asr corpus based on public domain audio books, *in* '2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)', IEEE, pp. 5206–5210.

Scheibler, R., Bezzam, E. & Dokmanic, I. (2018), 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, p. 351–355.

Schroeder, M. R. (1962), Natural sounding artificial reverberation, *in* 'Audio Engineering Society Convention 10', AES.

Srivastava, P., Deleforge, A. & Vincent, E. (2021), Blind room parameter estimation using multiple multichannel speech recordings, *in* '2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)', IEEE, pp. 226–230.

Szöke, I., Skácel, M., Mošner, L., Paliesek, J. & Černockỳ, J. (2019), 'Building and evaluation of a real room impulse response dataset', *IEEE Journal of Selected Topics in Signal Processing* **13**(4), 863–876.

Tang, Z., Bryan, N. J., Li, D., Langlois, T. R. & Manocha, D. (2020), 'Scene-aware audio rendering via deep acoustic analysis', *IEEE transactions on visualization and computer graphics* **26**(5), 1991–2001.

Wang, C., Jia, M., Li, M., Bao, C. & Jin, W. (2024*a*), Attention is all you need for blind room volume estimation, *in* 'ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 1341–1345.

Wang, C., Jia, M., Li, M., Bao, C. & Jin, W. (2024*b*), 'Exploring the power of pure attention mechanisms in blind room parameter estimation', *EURASIP Journal on Audio, Speech, and Music Processing* **2024**(1), 23.