ROOM ACOUSTIC MODELLING USING A HYBRID RAY-TRACING/FEEDBACK DELAY NETWORK METHOD

Haowen Zhao *

AudioLab
School of Physics, Engineering, and Technology
University of York
York, UK
haowen.zhao@york.ac.uk

Akihiko Suyama and Kazunobu Kondo

Research and Development Division
Yamaha Corporation
Hamamatsu, Japan
akihiko.suyama@music.yamaha.com
kazunobu.kondo@music.yamaha.com

Damian T. Murphy

AudioLab
School of Physics, Engineering, and Technology
University of York
York, UK
damian.murphy@york.ac.uk

ABSTRACT

Combining different room acoustic modelling methods could provide a better balance between perceptual plausibility and computational efficiency than using a single and potentially more computationally expensive model. In this work, a hybrid acoustic modelling system that integrates ray tracing (RT) with an advanced feedback delay network (FDN) is designed to generate perceptually plausible RIRs. A multiple stimuli with hidden reference and anchor (MUSHRA) test and a two-alternative-forced-choice (2AFC) discrimination task have been conducted to compare the proposed method against ground truth recordings and conventional RT-based approaches. The results show that the proposed system delivers robust performance in various scenarios, achieving highly plausible reverberation synthesis.

1. INTRODUCTION

Room acoustic simulation allows for the auralization of both real and virtual rooms. The real-time interactive nature of technologies such as gaming and virtual reality requires high computational efficiency to achieve real-time updates of room acoustics simulations in these scenarios [1][2]. In this background, the most commonly used modelling method is geometrical acoustics (GA)[3] with typical algorithms including ray tracing (RT) [4] and the image-source method (ISM) [5]. While ISM is quick to compute and accurate for low order reflections, the advantage in RT is that it can simultaneously simulate both specular and diffuse reflections, which are essential to achieve high-quality simulation [6]. However, accurately simulating diffuse reflections requires a large number of rays, which poses a challenge for real-time updates.

Artificial reverberation offers a computationally efficient alternative for simulating late reverberation, which is particularly suitable for real-time applications [7]. A commonly used algorithm for generating artificial reverberation with predefined multiband reverberation times is the feedback delay network (FDN) [7], which enables accurate control over room acoustic parameters such as reverberation time (T60).

A hybrid approach combining two methods in the time domain may offer a balance between perceptual plausibility and computational efficiency in acoustic modelling. Typically, the early reflections of a room impulse response (RIR) are computed using geometric methods such as the ISM and RT, while the reverberation tail is synthesized using artificial reverberation techniques such as the FDN. ISM is widely used in previous studies as the GA approach [8][9], as it can accurately simulate early reflections with minimal computational cost, particularly in simple room geometries. However, ISM typically only considers specular reflection and does not include scattering or diffraction effects [10]. Another key challenge is the uncertainty in determining whether the transition from specular to diffuse reflections occurs at a constant reflection order. In real rooms, this transition is more gradual, as scattering influences all diffuse reflections, including early non-specular reflections [11]. In contrast, although RT has a higher computational cost, it accounts for all types of reflection [3]. Previous studies have proposed running RT in parallel on GPUs, enabling greater efficiency and higher-order reflections [12]. This approach offers significant advantages in interactive environments [3].

Recent studies have introduced a range of optimizations for the FDN, which can be broadly categorized into two goals: generating smooth reverberation and achieving colorless reverberation. The former focuses on increasing echo density and ensuring a smooth reverberation buildup to avoid perceptually discrete or metallic artifacts. To this end, a filter feedback matrix (FFM) enabling the simulation of non-specular reflections and enhances the diffuseness and density of the reverberation [13]. The velvet feedback matrix (VFM), in particular, can produce high echo density reverberation with extremely low computational cost [13]. Similar optimization has also been achieved by directly employing velvet noise sequences and filters [14][15]. Meanwhile, colorless reverberation aims to reduce spectral coloration by optimizing the feedback matrix and gain structure. An optimized differentiable FDN has been developed to achieve this with fewer delay lines [16][17]. A recent study further integrates some of these strategies to establish a mapping between measured RIRs and FDN parameters [18], providing a key inspiration for this study.

This study introduces a hybrid reverberation modelling system that integrates scattering compensation and frequency-dependent decay control with a recently developed optimized FDN structure. The goal is to synthesize RIRs from low-order RT simulations that are perceptually similar to the measurements. The proposed system is structurally independent of the method used to generate the input RIR. RT was selected for the simulation in this study over

* This work was supported in part by Yamaha Corporation, Japan. The authors would also like to thank Ryo Tanaka for his support with this research.

the ISM due to its ability to handle both specular and diffuse reflections as well as higher orders. These properties are considered beneficial for the future development of more stable compensation strategies alghough such mechanisms are still under investigation. Multiple stimuli with hidden reference and anchor (MUSHRA) test and two-alternative-forced-choice (2AFC) listening tests were conducted to assess its subjective plausibility in multiple scenarios.

2. BACKGROUND

2.1. Ray tracing

RT is based on the assumption that sound propagates like rays, which is a valid approximation at mid to high frequencies, where the wavelength is small compared to room dimensions [7]. It enables the estimation of time-energy responses by tracing the paths of sound rays as they reflect off surfaces. Unlike the ISM, RT is a stochastic technique that uses Monte Carlo sampling to estimate the distribution of acoustic energy. This makes it especially useful for complex geometries and for modelling both specular and diffuse reflections [3]. Until now, there has been no real-time acoustic simulation software based solely on RT [19], unless it is a hybrid model such as room acoustics for virtual enviornments (RAVEN) [20], which combines the ISM, RT, and radiosity techniques.

2.2. Feedback delay network

A standard FDN consists of a set of delay lines interconnected via a feedback matrix defining each recirculating gain, with the transfer function in the z-domain as [21]:

$$H(z) = \frac{Y(z)}{X(z)} = \mathbf{c}^{\mathrm{T}} \left(\boldsymbol{D}_m(z^{-1}) - \boldsymbol{A}\right)^{-1} \boldsymbol{b} + d \tag{1}$$

where $X(z)$ and $Y(z)$ are the z-domain representations of input and output signals, respectively. $\boldsymbol{A}$ is an $N \times N$ feedback matrix that determines the energy coupling between the delay lines, and is typically chosen to be orthogonal to ensure energy preservation and desirable reverberation characteristics [22]. The matrix $\boldsymbol{D}_m(z) = \mathrm{diag}(z^{-m_1}, z^{-m_2}, \ldots, z^{-m_N})$ is a diagonal matrix representing the delay lengths, where each $\boldsymbol{m_i}$ corresponds to the length (in samples) of the $\boldsymbol{i}$-th delay line. The vectors $\boldsymbol{b}$ and $\boldsymbol{c}$ are of size $N \times 1$, defining the input and output gain distributions, respectively. The scalar term $\boldsymbol{d}$ represents the direct-path gain, allowing for a portion of the input to bypass the recursive structure.

2.3. Temporal transition strategies

A key consideration in hybrid modelling methods is how to manage the transition from early reflections to the late reverberant tail. Existing approaches differ in whether this transition is handled at a specific point in time, or distributed progressively over a period [23].

Many systems define a distinct crossover point at which the modelling method changes from ISM for early reflections, to statistical or artificial reverberators for the late reverberation. The crossover point is often determined using criteria, such as a fixed reflection order [10], a predefined amount of energy decay [8], or a perceptually motivated echo density profile [24][25]. While these approaches are straightforward to implement and allow for modular system design, they require careful calibration to ensure a perceptually smooth transition and to avoid artifacts such as spectral discontinuities or unnatural temporal textures.

Other approaches seek to reflect the continuous nature of the echo build-up more directly. Instead of switching models at a fixed time, these methods allow the contributions of specular and diffuse components to vary over time. Some redistribute reflection energy based on surface scattering coefficients [26], while others apply time-dependent envelopes to shape the emergence of the diffuse field [23]. Such strategies can improve perceptual smoothness and physical plausibility, albeit with greater model complexity.

2.4. Scattering effect in FDN

Traditional FDN designs face challenges in achieving high echo density while maintaining computational efficiency, particularly in low-dimensional systems where it is difficult to generate sufficiently dense impulse responses. To address this, the conventional scalar feedback matrix in the standard FDN can be replaced with a FFM composed of finite impulse response (FIR) or infinite impulse response (IIR) filters. This structure introduces time-spreading effects analogous to non-specular (scattering-like) reflections, thereby enabling the generation of significantly denser impulse responses while maintaining computational efficiency.

A particularly effective design is the VFM [13], inspired by the perceptual properties of velvet noise [27], a sparse sequence of ± 1 pulses that perceptually approximates white noise. By carefully controlling the pulse density δ in each FIR filter, the VFM achieves perceptually dense impulse responses at minimal computational cost. Specifically, the matrix is constructed so that each entry $A_{ij}(z)$ contains a small number of non-zero coefficients (i.e., pulses), and the overall structure remains paraunitary. For a VFM with K stages and $N \times N$ size, each matrix element becomes a sparse FIR filter of approximately $L = N^K$ taps, with only $\delta \cdot L$ non-zero elements.

2.5. Colorless FDN design

Another key challenge in the design of the FDN is to achieve colorless reverberation. This issue is particularly noticeable in low-dimensional FDN (i.e., $N = 4, 6, 8$), where the limited number of delay lines leads to a wide and uneven distribution of modal excitations, resulting in perceptually unnatural metallic ringing known as *coloration*.

A differentiable FDN optimization was introduced in [16], based on the finding that coloration is minimally affected by the choice of frequency-dependent attenuation [28]. In this approach, the feedback matrix A, as well as input and output gains b and c, are optimized using stochastic gradient descent method in the frequency domain, enabling the generation of perceptually colorless reverberation even with as few as 4 delay lines [16].

3. METHOD

The proposed system aims to synthesize perceptually plausible RIRs combining a low-order RT engine and an FDN to construct a combined RIR as shown in Figure 1. Firstly, a complete RT RIR is generated based on RT simulation. Then it is sent to estimate its frequency-dependent decay parameter via DecayFitnet [29]. DecayFitNet is a lightweight neural network to estimate the parameters of the multi-exponential decay into an amplitude, decay time, and a noise term. These estimates are used to shape the attenuation filters within a VFM FDN, producing a compensated RIR (VFM RIR) with improved echo density [13]. The RT RIR is also used to define the delay line configuration for the FDNs, which is passed to a differentiable FDN (Diff FDN). The VFM RIR is then analyzed a second time using DecayFitNet to obtain an updated decay parameter, which is then applied to Diff FDN for colorless optimization. The output DFDN RIR is the stable late reverberation tail for the specified geometry. To construct the final output, the mixing time is calculated from the DFDN RIR. The RT RIR is then truncated at this point, retaining only the early reflections. The early part is combined with the DFDN RIR to produce the final output system RIR.

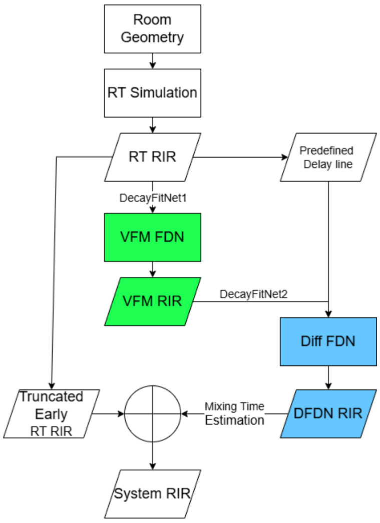


Figure 1: *Block diagram of the proposed hybrid reverberation system, with green indicating the first optimization stage and blue the second.*

3.1. Ray tracing system

A proprietary Yamaha developed RT simulation is used in this system and that prioritizes speed over accuracy for a 3D model. This sound ray simulation system is simplified for speed optimization. While it is capable of handling arbitrary 3D models, in this case, the room shape is also simulated as a cuboid. In this study, the RT simulation is configured with 30,000 rays and 10 reflection orders, which provides a computationally efficient approximation and is considered a low-order RT configuration in the context of hybrid reverberation synthesis.

The system adopts a modular design that the RT stage runs independently of the late reverberation synthesis. As a result, decay parameters are assumed to be unknown when the RIR is simulated by RT. Therefore, a complete RT RIR is used as the input for parameter estimation.

3.2. Scattering-like effect compensation

To compensate for the low echo density typically observed in the tail of RT RIRs, an FDN with a VFM is employed. The VFM is configured with $K = 4$ stages and a density factor of $\delta = 1$, forming a dense FIR feedback structure. While traditional VFM configurations use sparser impulses (e.g., $\delta = 1/30$) to reduce computation, this dense setup was selected to ensure a rapid increase in echo density and produce a highly diffuse reverberation tail.

The VFM structure directly influences the design of the delay line length in the FDN. Since VFM rapidly increases echo density, the selection of delay lengths becomes a vital factor in shaping the fine temporal structure of the synthesized late reverberation. In the proposed system, 8 delay lines were selected as mutually coprime intergers, uniformly distributed in order within a specified range based on the target room size (T60) and perceptual selection. First, a broad range for the delay line lengths was selected based on the target T60. Then, a finer range was chosen based on perceptual considerations: these intervals were found to best match the recorded RIRs in terms of echo density and the smoothness of late reverberation. For *M702* (T60 = 0.71s), a smaller set of prime numbered delay lengths was used in the range of 1000–2500 samples, while for the two larger rooms with T60 values around 1.0 s, longer primes were chosen in the range of 2500-5000 samples (the specific delay configurations for each room and detailed room information are listed in Section 4.1).

Before compensation, the complete RT RIR is analyzed using DecayFitNet to estimate its frequency-dependent decay envelope [29]. These estimates control a graphic equalizer (GEQ), which shapes the frequency-dependent decay envelope by adjusting the energy decay rate in each band individually [22].

3.3. Colorless FDN design

Despite compensation for echo density, some subtle high-frequency distinctness was still perceived in certain cases. This was because only 8 delay lines are included within the system to save computational cost. To further reduce spectral coloration and improve perceptual neutrality, the second stage of the system employs a colorless FDN architecture.

The implementation adopts the RIR2FDN framework [18], which includes the use of DecayFitNet to estimate an updated frequency-dependent T60, and an optimized differentiable FDN architecture introduced in [16]. Different from the first stage, these estimates from DecayFitNet are used to control a two-stage attenuation filter in [30], consisting of a shelving filter cascaded by a GEQ to achieve an accurate approximation of the frequency-dependent T60 throughout the frequency range. Compared to using a GEQ alone, the two-stage structure improves fitting accuracy especially at the lowest and Nyquist frequencies [30]. In this work, particular emphasis is placed on low-frequency control, as it plays a more perceptually significant role in achieving plausibility.

Particularly the DecayFitNet has been reused here to reanalyze the frequency-dependent T60. As noted in [13], a configuration with $K = 3$ for VFM compensation can cause deviations in the reverberation time of up to 20% compared to the ground truth. Since $K = 4$ is used in this system, the error introduced may be even more noticeable. Subjective listening and objective analysis confirmed that the T60 of the compensated IR systematically differ from the original RT RIRs. However, since the compensation parameters were perceptually selected to match the recorded RIRs, the resulting T60 aligns more closely with the recorded RIRs.

As this Diff FDN aims to achieve colorless effect based on VFM RIR, the delay line configuration used here is set the same as in the VFM FDN. In addition, an FFM is used within this framework, so it is also configured with $K = 4$ stages and a density factor of $\delta = 1$. For improved accuracy in shaping the frequency-dependent decay, a one-third-octave GEQ is employed within the two-stage attenuation filter design.

3.4. Transition Point

In this work a discrete transition point is used, selected on the basis of the estimated mixing time of the DFDN RIR. The mixing time is estimated using the echo density approach introduced by Abel and Huang [31]. In each frame, a Hann-windowed is used to compute the local energy standard deviation. The proportion of samples that exceed this threshold is normalized by the expected value of Gaussian noise. The first frame at which this normalized echo density reaches or exceeds one is defined as the mixing time.

At the estimated mixing time, the RT RIR is truncated, and the FDN-generated tail is appended. No crossfade or smoothing is applied, as the optimized FDN should well match the energy decay and echo density profile of the RT RIR, ensuring temporal and perceptual continuity despite the hard switch.

This choice is further supported based on [32]. In this work, it was shown that echo density evolution in FDNs aligns well with mixing time estimated based on Abel's method. As such, it offers perceptual plausibility and is therefore adopted in this work to define the crossover point between RT and FDN RIRs.

4. EVALUATION

To evaluate the perceptual performance of the proposed hybrid system, two listening tests were conducted. A MUSHRA-like test designed to evaluate the perceptual similarity of the proposed method compared to a real recorded reference firstly. The second was a 2AFC task aimed at evaluating the plausibility of the proposed method. Specifically, whether listeners could reliably distinguish its output from that of a real recording.

The test is run simultaneously in two places: AudioLab, University of York, UK and Yamaha Corporation, Hamamatsu, Japan. 24 people participated in the listening test in total, with 16 males and 8 females. They were between 18 and 60 years old, with a mean age of 30 years. None of them has any diagnosed hearing loss. They are all audio researchers or audio engineers. Sennheiser HD 650 headphones was used in York, while Yamaha HPH-MT8 studio monitor headphones was used in Hamamatsu [33]. Ethical approval was obtained from the Physical Science Ethics committee, University of York [34].

4.1. Stimuli

In this work, three anechoic sources are used: female speech, a cymbal, and an orchestra. All stimuli were 24-bit WAV files at a sampling rate of 48 kHz and normalized to a target amplitude level of -23.0 Loudness Units relative to Full Scale (LUFS). To reflect a range of realistic acoustic environments, three different rooms were selected for the listening tests, including a small conference room, a classroom, and a multi-purpose auditorium. All three spaces are real rooms located at the Yamaha Corporation headquarters in Hamamatsu, Japan, with the corresponding recorded RIRs. In all rooms, the source was placed in the center of the room, while the receiver was positioned at the center of one of the side walls. The RT simulation was conducted using the same combinations of sound source and receiver positions as in the RIR measurement. An overview of room information is provided in Table 1. Note that while the proposed system uses frequency-dependent T60 estimation and control, the values listed in Table 1 are frequency-averaged for brevity.

Table 1: Overview of the three acoustic environments used in the listening tests. The $T60$ was computed as the average across all frequency bands.

Room	$T60$	Description
M702	0.71 s	Small conference room with carpeted floor and soft furnishings.
Kenshu	1.08 s	Mid-sized classroom with reflective surfaces and partial acoustic treatment.
Shukai	1.01 s	Multipurpose auditorium with wooden floor and minimal absorption.

For the *M702*, the delay lengths were chosen in sample as $\mathbf{m}_{\text{M702}} = \{1009,\ 1217,\ 1429,\ 1657,\ 1847,\ 2069,\ 2281,\ 2477\}$, while for *Kenshu* and *Shukai*, the values were $\mathbf{m}_{\text{Kenshu, Shukai}} = \{2503,\ 2857,\ 3217,\ 3571,\ 3923,\ 4283,\ 4639,\ 4999\}$.

Five types of RIRs were used to generate the stimuli for the MUSHRA test:

- **Reference**: the recorded RIR measured in the corresponding real room.
- **Target**: the proposed hybrid system, which combines RT and FDN-based late reverberation.
- **RT-High**: a high-fidelity RT RIR computed using 30,000 rays and 10 reflection orders as in the system.
- **RT-Low**: a lower-fidelity RT RIR using 10,000 rays and 10 reflection orders.
- **Anchor**: a perceptual lower-bound RIR obtained by applying a 3 kHz low-pass filter to the reference.

These RIRs were then convolved with three anechoic sources, under three different room conditions (M702, Kenshu, Shukai), resulting in a total of 45 MUSHRA stimuli (3 rooms × 3 sources × 5 conditions).

In the 2AFC test, each stimulus pair consisted of the same anechoic source convolved with either a *Reference* or the *Target*. Each source–room combination was tested twice. Two additional control trials, identical to the training examples, were randomly inserted as hidden discrimination checks and excluded from the analysis, resulting in 20 trials in total (3 rooms × 3 sources × 2 repetitions + 2 additional control trials).

4.2. Procedure

The test consisted of two sessions presented in a fixed order: a MUSHRA-like quality rating test followed by a 2AFC plausibility task. Between the two sessions, participants participated in a 5-minute simple math game, which served as a distraction task to reduce the potential perceptual bias between sessions.

4.2.1. Listening test 1: MUSHRA

The MUSHRA test was conducted first and was based on the web audio API platform webMUSHRA [35]. Participants were allowed to adjust the system volume during this session to ensure comfortable listening levels. In each trial, participants were presented with five versions of the same source signal, each convolved with a different RIR as noted above. The five stimuli were presented in randomized order, with both the reference and the anchor included as hidden items.

Participants rated the perceived similarity of each sample to the hidden reference on a scale from 0 to 100. The playback was set to loop automatically within each trial, but the participants were free to pause as needed.

4.2.2. Listening test 2: 2AFC

The 2AFC test was conducted after the math game, as distraction task. The test was delivered via Qualtrics [36], a cloud-based platform that allows users to conduct online surveys for an online listening test. First, a brief training session was included. In each trial, participants were presented with two versions of the same sound source: one convolved with a *Reference* and the other with the *Target*. The two samples were presented in random order in each trial, and the trial sequence was designed to avoid consecutive trials featuring the same source material, to reduce cross-trial interference. They were asked to select the one they believed to be virtual.

Each combination of source and room was tested twice. In addition, two hidden control trials were randomly inserted to monitor the listener's attention. They were the same trials as the training examples (to which correct answers had been presented). Playback was not looped automatically, but participants could replay each pair as many times as needed before making a decision.

5. RESULTS

Among the 24 participants who completed both listening tests, data from 8 individuals were excluded from further analysis. 3 of them were due to their failure to correctly identify both hidden discrimination check trials in the 2AFC test, and 5 of them had a hard time recognizing the reference in more than 15% of the trials. As a result, the reported findings are based on the remaining 16 participants.

5.1. MUSHRA

Results of the MUSHRA test are shown in Figure 2. Figure 2(a) shows the overall score distribution across all sources and rooms. As the normality assumption was not met based on the Shapiro-Wilk test, the Friedman test was applied to analyze system ratings across room and source conditions. For conditions showing significance, pairwise Wilcoxon signed-rank tests were conducted, and multiple comparisons were corrected using the false discovery rate (FDR) method.

Compared with *RT-high*, *Target* significantly outperformed it in 7 out of 9 conditions. Among the remaining two trials, *ShukaiCymbal* showed borderline significance ($p = 0.063$), while *KenshuOrchestra* showed a non-significant difference despite a higher mean rating ($p = 0.124$). In contrast, in all 9 conditions, *Target* received significantly lower ratings in all trials comparing to *Reference* ($p < 0.001, FDR - adjusted$), consistently showing the existence of a perceptual difference.

A breakdown by room showed a significant improvements in *M702* ($p = 0.030$) and *Kenshu* ($p = 0.018$), with a borderline result in *Shukai* ($p = 0.067$). When grouped by source, *Target* showed a significant advantage in *Speech* ($p = 0.013$), while the differences for *Orchestra* and *Symbal* were not statistically significant ($p = 0.374$ for both). Particularly, *Target* performed best in *M702Speech*, achieving a mean rating of 83.44, which was the closest to the *Reference* among all tested conditions.

5.2. 2AFC

The analysis covers 18 randomized questions per participant (excluding two hidden discrimination tasks) as presented in Table 2. The mean accuracy across all trials was 50.31%, not significantly different from the 50% guessing level ($p = 0.4777$). In 2AFC plausibility tests, since no explicit reference was provided, this is considered sufficient to indicate that the *System* is perceptually highly plausible comparing with the *Reference* [37]. Signal detection analysis was performed with a sensitivity index of $d' = 0.0158$, which falls within the range typically interpreted as low discriminability in forced-choice paradigms [38]. The proportion of first-answer 'A' responses was 54.69%, with a bias z-score of +1.6771, showing a slight but statistically non-significant tendency to choose 'A'.

Table 2: Signal Detection and Plausibility Test Results

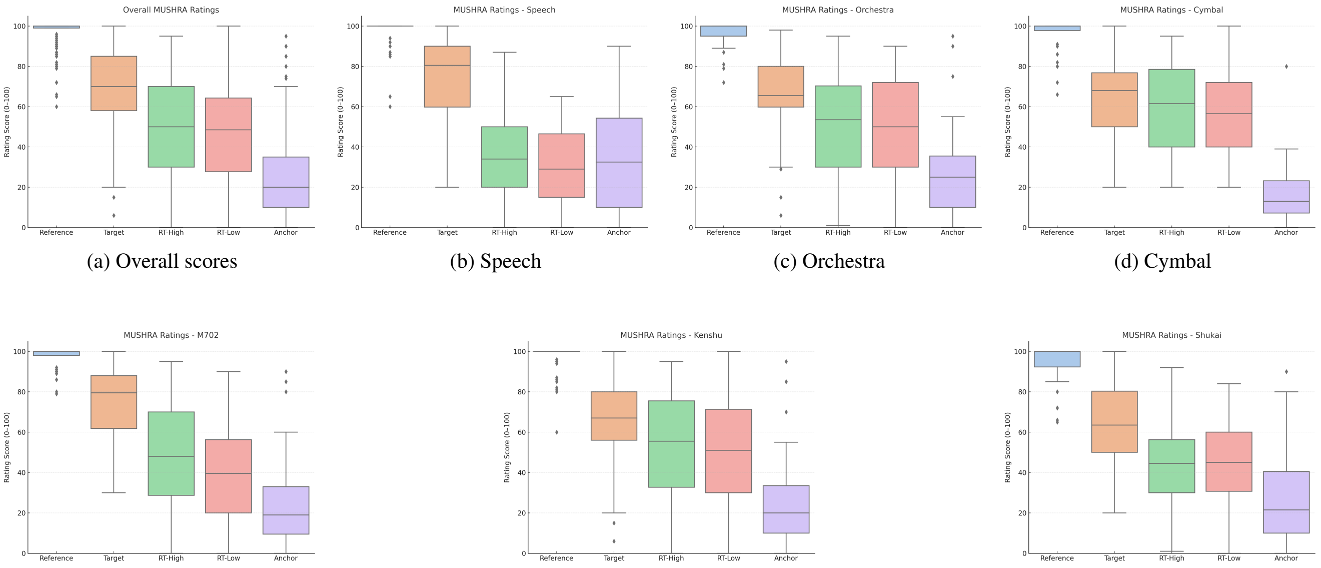
Metric	Value
Mean Accuracy	50.31%
Sensitivity d'	0.0158
P(Choose A)	54.69%
Bias (z-score of $P(A)$)	$+1.6771$
Binomial Test (p-value)	0.4777

Figure 3 shows the distribution of the accuracy of the participants. The results are mainly clustered around 50%, with no outliers, suggesting consistency between listeners. Figure 4, shows the average accuracy by *Room* and *Source*. Most conditions resulted in performance near the guessing rate. All combinations with cymbal sounds showed slightly higher accuracy than average, especially in M702. However, the rest of combinations in *Kenshu* achieved lower accuracy than average, this may be because flutter echoes in the Kenshu had a greater impact on non-transient sources like speech and orchestra, which lack strong high-frequency content to mask such artifacts.

6. DISCUSSION

The results of the listening test suggest that the proposed hybrid system achieves a generally plausible reverberation quality in several acoustic scenarios. In the MUSHRA test, the system outperformed both high- and low-order RT baselines, especially for speech stimuli. In the 2AFC test, the overall results showed limited discriminability, with listener accuracy largely clustered around the chance level across different room and source conditions.

The choice of sound source influenced the test result. In the 2AFC test, *Cymbal* consistently scored above the chance level (50%), which indicates that high-frequency transients of the cymbal can enhance perceptual discriminability in short comparison tasks, particularly when overall plausibility remains limited. In contrast, *Orchestra* showed increased 2AFC accuracy with larger room sizes, possibly because spectral masking in the small room increased the perceived plausibility of the simulation, even with broadband content.



(a) Overall scores (b) Speech (c) Orchestra (d) Cymbal

(e) M702 (f) Kenshu (g) Shukai

Figure 2: MUSHRA rating distributions across system conditions. (a) Overall scores, (b–d) split by audio source type, and (e–g) by room condition.

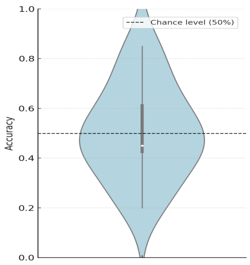


Figure 3: Participant accuracy distribution in the 2AFC task. The dashed line indicates the 50% guessing level.

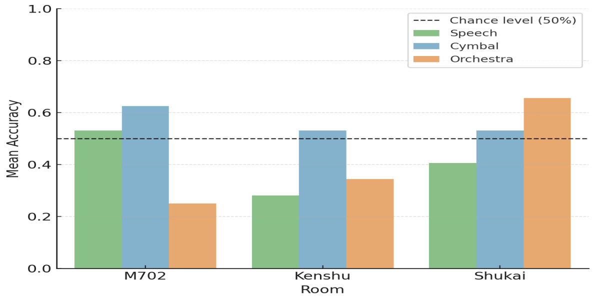
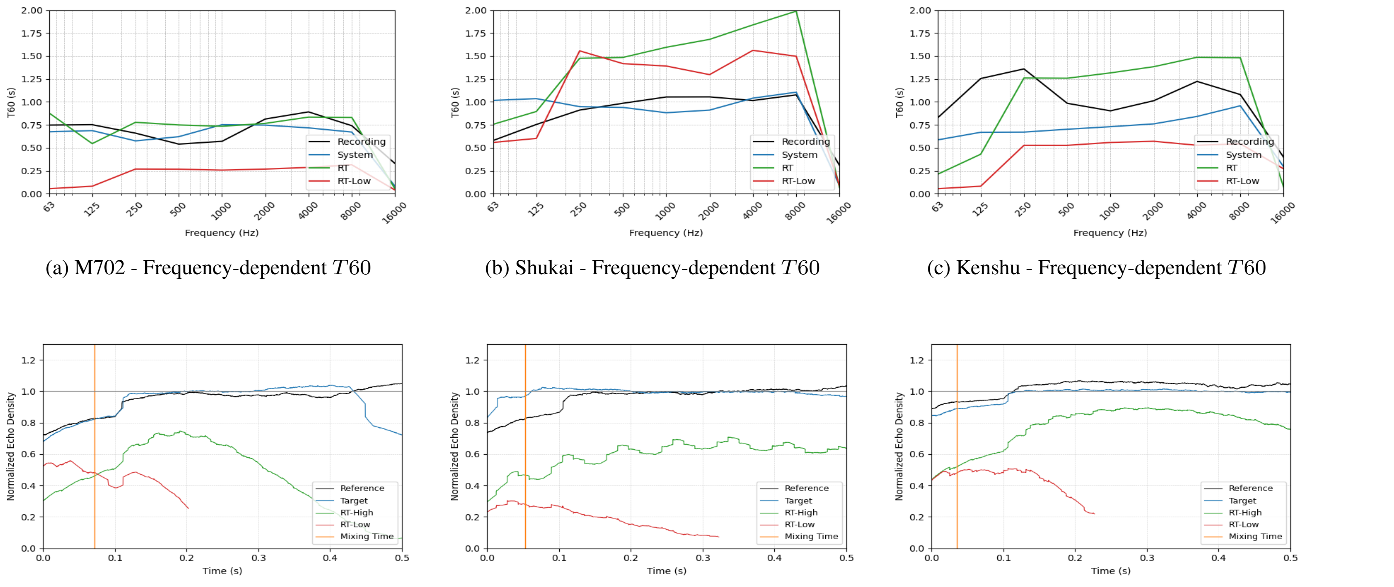


Figure 4: Mean recognition accuracy grouped by room and source. Bars represent average performance, with the dashed line indicating chance level.

The room-dependent deviations in T60 also have a great influence as also shown in Figure 5. While *M702* showed good alignment with *Reference*, *Shukai* and *Kenshu* revealed mismatches, especially in *Kenshu*. This highlights a key limitation of the current system: the echo density compensation achieved by the VFM has not yet been reliably connected to the decay times. This mismatch had a notable impact on the MUSHRA results. As shown in Figure 5, the *Target* in *Kenshu* received ratings that were much closer to the RT baselines compared to the other two rooms, and deviated more noticeably from *Reference*.

However, in the 2AFC test, the *Kenshu* condition has the best performance. This suggests that although perceptual differences exist between *Target* and *Reference*, participants still perceived *Target* with similar echo density as highly plausible where no explicit reference was available. Although both 2AFC and MUSHRA were used to evaluate the perceptual quality of the proposed system, their results reflect distinct but complementary perceptual criteria. The 2AFC task emphasizes plausibility, whether a result sounds plausible and acceptable solo, without requiring listeners to compare it to a reference directly. In contrast, the MUSHRA test makes a direct comparison among multiple systems using a fixed reference, making it more sensitive to subtle spectral or temporal deviations. As a result, a system may be judged plausible in 2AFC but still receive lower MUSHRA ratings due to small perceptual differences. Similar findings were reported in [39], who showed that plausibility and similarity ratings can be different when evaluating virtual acoustic environments. These results suggest that combining both methods provides a more comprehensive understanding of perceptual outcomes, with 2AFC capturing plausibility thresholds and MUSHRA showing the relative perceptual similarity between systems.

A possible limitation from the RT is the relatively short duration of the synthesized RIRs, which typically end around 0.5 seconds. This reflects the limited energy decay captured by low-order RT and results in a reduced dynamic range for late reverberation analysis. While such constraints may affect the accuracy of T60



(a) M702 - Frequency-dependent $T60$

(b) Shukai - Frequency-dependent $T60$

(c) Kenshu - Frequency-dependent $T60$

(d) M702 - Echo Density

(e) Shukai - Echo Density

(f) Kenshu - Echo Density

Figure 5: Comparison of reverberation characteristics across rooms. Top row: Frequency-dependent $T60$ plots of the recording, proposed system, and RT baselines. Bottom row: Echo density curves showing temporal build-up behavior across models.

estimation and limit perceptual plausibility, the regression regions used for decay time calculation remained within a valid dynamic range in all tested conditions. Furthermore, as the proposed system builds based on the RT RIRs, their lengths were kept consistent to ensure fair comparison. Extending the duration of synthesized RIRs may further enhance both acoustic accuracy and perceptual plausibility in future implementations.

7. CONCLUSION

This paper introduced a hybrid reverberation system that combines low-order RT and FDN. To address the lack of scattering in RT RIRs, a VFM was employed to enhance echo density in the early response. Subsequently, a recently proposed mapping from RIRs to FDN parameters was used to derive the structure of the final FDN, ensuring colorless and smoothly decaying reverberation. The modular framework also allows for extensions toward real-time reverberation synthesis, in which early reflections could be computed on the fly and combined with the pre-optimized FDN-based late reverberation tail. The proposed system narrows the gap between low-cost acoustic simulation and perceptually plausible reverberation, providing a controllable synthesis framework that brings new possibilities for designing RIRs for virtual rooms.

Listening tests, including MUSHRA and 2AFC, demonstrated that the system achieves perceptually plausible results under various conditions, in some cases approaching the quality of real recordings. Future work will focus on improving the scattering compensation strategy by making the design of the VFM more physically informed.

8. REFERENCES

[1] Carl Schissler, Aaron Nicholls, and Ravish Mehra, “Efficient hrtf-based spatial audio for area and volumetric sources,” *IEEE Trans. Visualization and Computer Graphics*, vol. 22, no. 4, pp. 1356–1366, 2016.

[2] Michael Vorländer, *Auralization*, Springer, 2020.

[3] Lauri Savioja and U Peter Svensson, “Overview of geometrical room acoustic modeling techniques,” *J. of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015.

[4] Anna Pompei, MA Sumbatyan, and NF Todorov, “Computer models in room acoustics: The ray tracing method and the auralization algorithms,” *Acoustical Physics*, vol. 55, pp. 821–831, 2009.

[5] Marc Aretz, Pascal Dietrich, and Michael Vorländer, “Application of the mirror source method for low frequency sound prediction in rectangular rooms,” *Acta Acustica united with Acustica*, vol. 100, no. 2, pp. 306–319, 2014.

[6] Bengt-Inge Dalenbäck, Mendel Kleiner, and Peter Svensson, “A macroscopic view of diffuse reflection,” *J. of the Audio Engineering Society*, vol. 42, no. 10, pp. 793–807, 1994.

[7] Vesa Valimaki, Julian D Parker, Lauri Savioja, Julius O Smith, and Jonathan S Abel, “Fifty years of artificial reverberation,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, 2012.

[8] Pavel Zahorik, “Perceptually relevant parameters for virtual listening simulation of small room acoustics,” *J. of the Acoustical Society of America*, vol. 126, no. 2, pp. 776–791, 2009.

[9] Marc Aretz, René Nöthen, Michael Vorländer, and Dirk Schröder, “Combined broadband impulse responses using

fem and hybrid ray-based methods," in *EAA Symp. on Auralization*, 2009, pp. 15–17.

[10] Torben Wendt, Steven Van De Par, and Stephan D Ewert, "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," *J. of the Audio Engineering Society*, vol. 62, no. 11, pp. 748–766, 2014.

[11] Stephan D Ewert, Nico Gößling, Oliver Buttler, Steven van de Par, and Hongmei Hu, "Computationally-efficient rendering of diffuse reflections for geometrical acoustics based room simulation," *Acta Acustica*, vol. 9, pp. 9, 2025.

[12] Lauri Savioja, Dinesh Manocha, and M Lin, "Use of gpus in room acoustic modeling and auralization," in *Proc. Int. Symp. on Room Acoustics*, 2010, p. 3.

[13] Sebastian J Schlecht and Emanuël AP Habets, "Scattering in feedback delay networks," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 28, pp. 1915–1924, 2020.

[14] Jon Fagerström, Benoit Alary, Sebastian Schlecht, and Vesa Välimäki, "Velvet-noise feedback delay network," in *Int. Conf. on Digital Audio Effects (DAFx)*, 2020, pp. 219–226.

[15] Vesa Välimäki and Karolina Prawda, "Late-reverberation synthesis using interleaved velvet-noise sequences," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 29, pp. 1149–1160, 2021.

[16] Gloria Dal Santo, Karolina Prawda, Sebastian Schlecht, and Vesa Välimäki, "Differentiable feedback delay network for colorless reverberation," in *Int. Conf. on Digital Audio Effects (DAFx)*, 2023, pp. 244–251.

[17] Gloria Dal Santo, Karolina Prawda, Sebastian J Schlecht, and Vesa Välimäki, "Feedback delay network optimization," *arXiv e-prints*, pp. arXiv–2402, 2024.

[18] Gloria Dal Santo, Benoit Alary, Karolina Prawda, Sebastian Schlecht, and Vesa Välimäki, "Rir2fdn: An improved room impulse response analysis and synthesis," in *Int. Conf. on Digital Audio Effects (DAFx)*, 2024, pp. 230–237.

[19] Hanna Autio, *Room Acoustic Ray-Tracing: Understanding, evaluating and expanding the toolset*, Lund University, 2024.

[20] Dirk Schröder and Michael Vorländer, "Raven: A real-time framework for the auralization of interactive virtual environments," in *Forum Acusticum*. Aalborg Denmark, 2011, pp. 1541–1546.

[21] Jean-Marc Jot and Antoine Chaigne, "Digital delay networks for designing artificial reverberators," in *Audio Engineering Society Convention 90*. Audio Engineering Society, 1991.

[22] Sebastian Schlecht, "Fdntb: The feedback delay network toolbox," in *Int. Conf. on Digital Audio Effects (DAFx)*. DAFx, 2020, pp. 211–218.

[23] Wouter Wittebol, Huiqing Wang, Maarten Hornikx, and Paul Calamia, "A hybrid room acoustic modeling approach combining image source, acoustic diffusion equation, and time-domain discontinuous galerkin methods," *Applied Acoustics*, vol. 223, pp. 110068, 2024.

[24] Eric A Lehmann and Anders M Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1429–1439, 2009.

[25] Alexander Lindau, Linda Kosanke, and Stefan Weinzierl, "Perceptual evaluation of model-and signal-based predictors of the mixing time in binaural room impulse responses," *J. of the Audio Engineering Society*, vol. 60, no. 11, pp. 887–898, 2012.

[26] Stephan D Ewert, Nico Gößling, Oliver Buttler, Steven van de Par, and Hongmei Hu, "Computationally-efficient and perceptually-motivated rendering of diffuse reflections in room acoustics simulation," *arXiv preprint arXiv:2306.16696*, 2023.

[27] Vesa Välimäki, Bo Holm-Rasmussen, Benoit Alary, and Heidi-Maria Lehtonen, "Late reverberation synthesis using filtered velvet noise," *Applied Sciences*, vol. 7, no. 5, pp. 483, 2017.

[28] Janis Heldmann and Sebastian J Schlecht, "The role of modal excitation in colorless reverberation," in *Int. Conf. on Digital Audio Effects (DAFx)*, 2021, pp. 206–213.

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

[30] Vesa Välimäki, Karolina Prawda, and Sebastian J Schlecht, "Two-stage attenuation filter for artificial reverberation," *IEEE Signal Processing Letters*, vol. 31, pp. 391–395, 2024.

[31] Jonathan S Abel and Patty Huang, "A simple, robust measure of reverberation echo density," in *Audio Engineering Society Convention 121*. Audio Engineering Society, 2006.

[32] Sebastian J Schlecht and Emanuël AP Habets, "Feedback delay networks: Echo density and mixing time," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 374–383, 2016.

[33] Yamaha Corporation, "HPH-MT8 Headphones," `https://uk.yamaha.com/en/products/proaudio/headphones/hph-mt8/index.html`, 2025, Accessed: March 25, 2025.

[34] University of York, "Academic ethics and compliance committee," 2025, Accessed: 25-March-2025.

[35] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre, "webmushra—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[36] "Qualtrics," Qualtrics, 2025, Accessed: March 2025.

[37] Alexander Lindau and Stefan Weinzierl, "Assessing the plausibility of virtual acoustic environments," *Acta Acustica united with Acustica*, vol. 98, no. 5, pp. 804–810, 2012.

[38] Anamaria Madalina Nastasa, Nils Meyer-Kahlen, and Sebastian Schlecht, "Assessing room acoustic memory using a yes/no and a 2-afc paradigm," in *Nordic Sound and Music Conference*. Aalborg Universitet, 2021, pp. 61–65.

[39] Stefan Fichna, Steven van de Par, and Stephan D Ewert, "Evaluation of virtual acoustic environments with different acoustic level of detail," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 2023, pp. 1–6.