Audio Engineering Society

Conference Paper

Presented at the AES 2026 International Conference on
Audio for Virtual and Augmented Reality and Immersive Games
2026 June 30 – July 3, Paris, France

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.

Assessing Efficient Auralization Methods in Architectural Virtual Environments

Georgios Papadimitriou[1, 2], Nils Meyer-Kahlen[3, 4], Hanna Järveläinen[2], Achilleas Xydis[1], Carlotta Daro[1], Matthias Kohler[1], and Rama Gottfried[2]

[1]*Gramazio Kohler Research, ETH Zurich, Switzerland*
[2]*Institute for Computer Music and Sound Technology, Zurich University of the Arts, Switzerland*
[3]*Dpt. of Information and Communications Engineering, Aalto University, Finland*
[4]*Dyson School of Design Engineering, Imperial College London, UK*

Correspondence should be addressed to Georgios Papadimitriou (papadimitriou@arch.ethz.ch)

ABSTRACT

Auralization enables multisensory evaluation of architectural designs in Virtual Reality (VR), yet physically accurate acoustic simulations remain computationally prohibitive for interactive workflows. This study investigates efficient artificial reverberation methods as lightweight proxies for different stages of VR-based architectural design. After assessing the predictive capabilities of geometrically informed models, a hierarchical 3-Alternative-Forced-Choice listening experiment with a transferring task paradigm was conducted in VR using binaural audio. In this experiment, the measured room impulse responses of a physical space in untreated and acoustically treated conditions were compared with those from five auralization techniques. These techniques ranged from industry-standard simulations to artificial reverberators. An initial investigation revealed that neither of the artificial reverberation methods that could take material parameters as input did not provide credible results. Thereafter, all methods were calibrated to the Energy Decay Curves of reference simulations. The results revealed that the pure Image-Source Method was easily detected, likely because the late reverberation's temporal density was insufficient. When incorporating a dense late reverberant tail, no credible difference was detected at this sample size between efficient methods and high-fidelity simulations, although directional trends were observed. Participants prioritized the timbral quality of late reverberation over geometrically accurate early reflections. This suggests that computationally efficient models can serve as convincing, scalable rendering tools for interactive design and presents this audiovisual VR paradigm as an ecologically valid platform for multisensory architectural assessment.

1 Introduction

Virtual Reality (VR) is increasingly adopted in architectural design workflows, offering an immersive medium for exploring unbuilt spaces [1]. By enabling the assessment of acoustic qualities [2], integrating auralization extends this capability into a powerful multisensory tool. However, generating perceptually adequate reverberation remains constrained by processing resources. Although wave-based acoustics provide rigorous reconstructions, their computational requirements limit their usage [3]. Consequently, Geometrical Acoustics (GA)

methods using the Image-Source Method (ISM) [4, 5] and Ray Tracing (RT) [6] are still the industry standard [7] in architectural acoustics. The practical application of acoustic simulations for rapid, VR-based, iterative architectural design is limited by the constraints imposed by required model simplifications, the need to continually update them, and the computational demands of high-order reflections required for accuracy.

Overcoming these limitations requires a compromise between computational complexity and perceptual fidelity. This trade-off drives the development of efficient artificial reverberation algorithms, including Directional Feedback Delay Networks (DFDN) [8], geometrically informed Scattering Delay Networks (SDN) [9], and hybrid models that pair geometric early reflections with artificial late reverberation [10]. The underlying idea is that precise geometric replication of complex architectural spaces can be bypassed if the reverberator provides sufficient temporal density and an accurate spectral envelope, which compensates for spatial simplification in ambisonic reverberation [11].

Efficient artificial reverberation methods could be integrated into VR-based architectural workflows at two distinct stages: first, geometrically informed reverberators that take material absorption and room volume as input could serve as proxies for simulators, allowing the acoustics of unbuilt designs to be auralized without running a full GA simulation; second, they could be used merely as real-time rendering proxies. This approach can facilitate 3- or 6-Degree-of-Freedom (3-, 6-DoF) spatial audio by avoiding storage and loading of dense grids of room impulse responses (RIRs) for convolution-based rendering. Even here, there may be different use cases: artificial models could be tuned to match in-situ measurements to virtually replicate an existing space, or they could be tuned to the results of a GA simulation when designing a new space or an acoustical intervention.

Despite practical need, a gap remains in the perceptual validation of these methods in applied architectural contexts. While round-robin studies have evaluated various simulators [12], few have simultaneously compared artificial reverberation methods with measured room impulse responses (RIRs) and industry-standard simulations [13]. For many VR-based applications, strict physical authenticity is not required, provided that the auralization is functional and convincing [14]. Moreover, spatial evaluation is inherently multimodal; visual representation fundamentally contextualizes the acoustic experience [15]. Consequently, an ecologically valid approach is needed using an audiovisual testing paradigm rather than a traditional "audio-only" listening test.

This study evaluates the perceptual quality of lightweight, code-based reverberation algorithms that avoid computationally expensive geometric simulations for architectural auralization, with the goal of integrating them into interactive design workflows. We first test whether geometrically informed models could serve as proxies for GA-based simulations. After observing significant errors in some models, we transition to evaluating them only for perceptually valid rendering. To ensure that listeners assessed spatial and temporal texture rather than energetic deviations, all methods were calibrated to match reference Energy Decay Curves (EDC) at this stage. For the calibrated reverberators, we present a hierarchical audiovisual listening experiment testing four methods (ISM, SDN, DFDN, and a Hybrid model) in two different acoustic conditions of the same physical space, measured before and after acoustic treatment. To target applications in diverse architectural workflows, the assessment spans three comparative phases, in which all rendering methods are compared against a different reference: (a) measured RIRs, (b) GA simulations, and (c) hybrid RIRs (which was the best performing model in the previous stages). To ensure ecological validity, the participant pool consisted mainly of practicing architects and architectural researchers who evaluated acoustics within a congruent virtual reality environment.

2 Simulation and Rendering Methods

The study uses a shoebox office environment measuring $13.5 \times 9.6 \times 4.8$ m (622 m^3). To evaluate different acoustic contexts, the room was captured under two conditions: untreated/reverberant and treated/absorptive, using a Sennheiser Ambeo VR microphone (Sennheiser, Wedemark, Germany) and an omnidirectional DS3 Dodecahedron speaker (NTi Audio AG, Schaan, Liechtenstein).

Five different reverberation rendering techniques were used to generate RIRs, as described below. All simulated RIRs were rendered in first-order Ambisonics to maintain a constant spatial resolution. This spatial format enables dynamic binaural decoding and allows participants to use natural 3-Degree-of-Freedom (3-DoF)

head movements to evaluate the reverberant environment ecologically. Wave-based solvers were excluded from the study due to prohibitive computational costs that make them inefficient for real-time calculations.

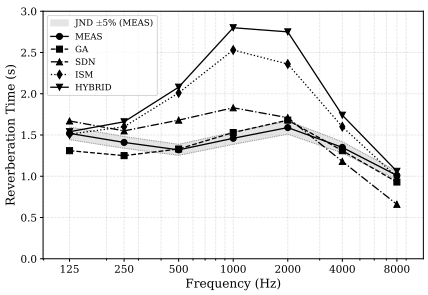
Geometrical acoustics (GA): Acoustic simulation using *Odeon Room Acoustics Software* which is widely accepted as an industry standard [16].

Image Source Method (ISM): A simplified geometric method that utilizes *Pyroomacoustics*, an open-source implementation of image source and ray tracing [17]. The model was restricted to pure ISM (omitting ray tracing) with order set to 100.

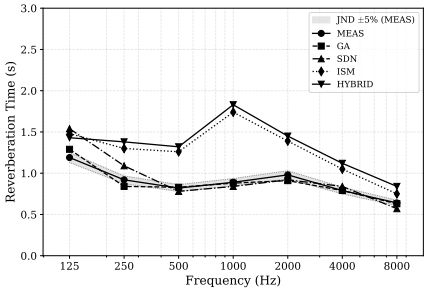
Hybrid Model: A composite method adapted for this study that synthesizes a first-order Ambisonics RIR by combining a physically simulated early part with a Feedback Delay Network (FDN) for the late reverberation. Early reflections are calculated using the *Pyroomacoustics* implementation of the ISM [17]. For the reverberant tail, the Ambisonic channels are analyzed separately using *DecayFitNet* [18] to estimate the frequency-dependent reverberation time and energy levels across eight octave bands (63Hz–8 kHz). Based on these extracted parameters, the late tail is synthesized using the *RIR2FDN* method [19]. The two segments are combined at the mixing time, estimated using the echo density method [20].

Scattering Delay Network (SDN): An approach that balances geometric validity with delay network efficiency [9] by simulating first-order wall reflections precisely while generating the reverberant tail through recursive scattering nodes. The SDN VST implementation by Fontana et al. [21] is used to synthesize RIRs. The air-absorption setting is activated, despite having been shown to affect the resulting reverberation time in unexpected ways [13].

Directional Feedback Delay Network (DFDN): A purely artificial reverberator, custom implemented in Python for this study based on the theoretical framework proposed by Alary et al. [8]. Here, 32 delay lines are used in eight groups, each carrying a full set of four first-order Ambisonic channels (ACN/SN3D). Source directionality is encoded via first-order spherical harmonic coefficients. The feedback matrix follows the Kronecker product structure, with an 8×8 random orthogonal matrix and scalar absorption gains controlling the decay.



(a) T_{30} in the untreated ("Wet") condition before calibration.



(b) T_{30} in the treated ("Dry") condition before calibration.

Fig. 1: Initial reverberation times T_{30} (s) of auralization methods with initial material absorption coefficients as input.

2.1 Results for Uncalibrated Reverberators

To assess whether simple, geometrically informed methods (ISM, SDN, Hybrid) could be used to predict acoustics, they were initialized using the same absorption coefficients as in the reference GA simulation (α) across seven octave bands (125Hz–8kHz). The untreated ("Wet") condition utilized: *Suspended ceiling* α = [0.30, 0.45, 0.52, 0.58, 0.40, 0.30, 0.25], *Plasterboard:* α = [0.20, 0.20, 0.16, 0.11, 0.12, 0.20, 0.20], *Concrete:* α = [0.01, 0.01, 0.02, 0.02, 0.02, 0.05, 0.05], and *Glass window:* α = [0.10, 0.07, 0.05, 0.03, 0.02, 0.02, 0.02]. For the treated ("Dry") condition, an acoustic material with absorption coefficients α = [0.20, 0.57, 0.86, 0.77, 0.70, 0.70, 0.65] was applied.

Fig. 1 shows that the GA simulation yielded reverberation times T_{30} that closely matched the measured reverberation time using these physical parameters. However, the real-time approximations (ISM, SDN, and hybrid) varied considerably. Only the ISM in the dry con-

dition remained within reasonable bounds at mid- and high-frequency bands. Even though we did not expect perfect results and believe that interactive auralization cannot strive for authenticity, such reverberation-time errors would be clearly detected. Thus, we can already conclude that the predictive capabilities of these simple models are still limited. However, they may be useful at the rendering stage. Therefore, the raw real-time approximations must be calibrated to match the targeted reverberation time of the measured RIRs.

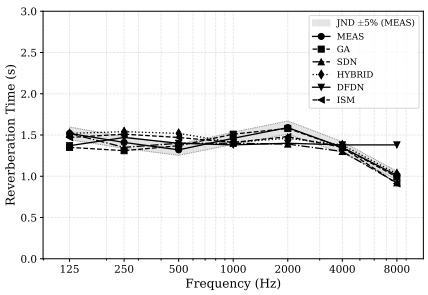
2.2 Calibration for Perceptual Rendering

To ensure that the listening test evaluated properties beyond the obvious decay discrepancies, an iterative manual tuning phase aligned the simulated Energy Decay Curves (EDC) with the measured target. Geometric methods achieved this through adjusted absorption coefficients. At this stage, the non-geometry-based DFDN can also be included. Although exhibiting deviations in extreme frequency bands (125Hz and 8kHz), which are likely implementation artifacts rather than fundamental limits, this custom DFDN achieves adequate mid-frequency accuracy (500Hz–2kHz). Given that the experimental stimuli consisted of anechoic speech signals, which are dominated by mid-frequency energy, these discrepancies at the spectral extremes have a negligible impact on the overall perceptual evaluation. After calibration, most methods' T_{30} converged within or near the assumed Just Noticeable Difference (JND) threshold of 5% [22] in the perceptually critical mid-frequencies (500Hz–1kHz), see Fig. 2.

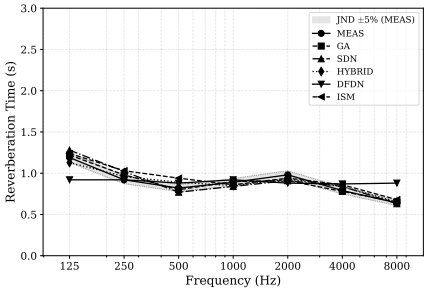
3 Perceptual Experiment

3.1 Participants

The experiment included 22 participants (age 25–58 y, mean = 37.4 y, SD = 9.4 y). Prior to data collection, participants self-reported their experience in acoustics and architecture on a scale between 1 (Novice) and 5 (Expert). The majority reported limited experience in acoustics and audio engineering (median = 2), while most reported advanced architectural knowledge (median = 4). All participants reported normal hearing and normal or corrected-to-normal vision and gave informed consent to participate in the experiment and to publish the results.



(a) T_{30} in the untreated ("Wet") condition after calibration.



(b) T_{30} in the treated ("Dry") condition after calibration.

Fig. 2: Calibrated reverberation times T_{30} (s) of the auralizations.

3.2 Experimental Design

To evaluate the utility of real-time algorithms, we conducted an experiment to predict detection rates and ratings using two crossed predictors: auralization method and acoustic condition (untreated/wet and treated/dry). The hierarchical strategy was structured into three phases, each representing a distinct architectural use case. Phase 1 compared all rendering methods (GA, Hybrid, ISM, SDN, and DFDN) with measured RIRs, mirroring the calibration of existing spaces. Phase 2 used the full-scale GA simulation as a benchmark against faster approximations (Hybrid, ISM, SDN, DFDN) to assess their viability for more efficiently auralization of predictions for unbuilt environments. Phase 3 evaluated the most computationally efficient algorithms (ISM, SDN, DFDN) against the high-performing Hybrid model for interactive real-time VR rendering. To minimize order effects, trials across all three conceptually distinct phases were fully randomized.

Training (5 Trials): Participants underwent a training phase with distinct acoustic differences to familiarize themselves with the task and interface (these trials were excluded from the analysis).

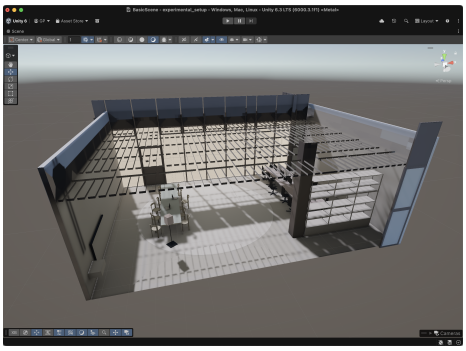
Phase 1 (20 Trials): All simulation methods were compared with the measured RIR to establish a baseline with respect to the physical world (5 Methods $\times$ 2 Conditions $\times$ 2 Reps).

Phase 2 (16 Trials): Open-source approximations (Hybrid, ISM, SDN, DFDN) were compared against the GA benchmark to test if fast algorithms can serve as valid auralization proxies for industry-standard software (4 Methods $\times$ 2 Conditions $\times$ 2 Reps).

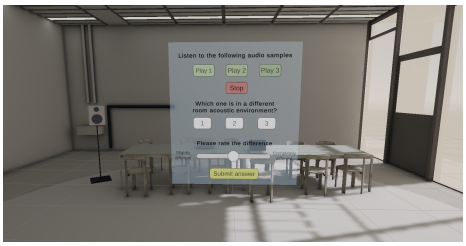
Phase 3 (12 Trials): Efficient methods (ISM, SDN, DFDN) were compared against the Hybrid model, chosen as a robust baseline based on pilot testing, to assess the perceptual cost of algorithmic simplification (3 Methods $\times$ 2 Conditions $\times$ 2 Reps).

3.3 Visual Rendering

With VR-based architectural design application in mind, and drawing on previous research showing that congruent visuals stabilize spatial perception without introducing undue acoustic bias [23], the 3D model was used to provide a multimodal framework for participants. The visual environment, depicted in Fig. 3a, was presented as a high-quality architectural abstraction and was maintained under both acoustic conditions to align with typical early-stage design aesthetics and simulate rapid auralization before architectural updates. In addition, a virtual loudspeaker was positioned at the exact coordinates of the RIR measurement source to provide a spatial reference for auralized audio. The virtual environment was constructed from a detailed 3D model of the physical office created in Rhinoceros 3D (Robert McNeel & Associates, TLM Inc.) imported into Unity (Unity Software Inc., San Francisco, US), and presented using a Meta Quest 3 Head-Mounted Display (HMD) (Meta Platforms, Inc., Menlo Park, California, US). Note that while intended as a neutral abstraction, the model's plain surfaces might visually resemble a reverberant space, introducing an inherent audiovisual incongruence in the "Dry" condition.



(a) 3D model of the office in Unity.



(b) First-person view of the Graphic User Interface (GUI).

Fig. 3: Visual rendering of the office in the experiment.

3.4 Audio Playback

Audio playback was decoupled from Unity, which acted solely as a control interface, transmitting Open Sound Control (OSC) messages to the Digital Audio Workstation "Reaper" (Cockos Inc., San Francisco, US). Within Reaper, Ambisonic RIRs were convolved with anechoic speech samples [24] and decoded to binaural audio using the SPARTA plugin suite. Real-time head orientation data (yaw, pitch, roll) was streamed via OSC from Unity to the binaural decoder to enable 3-Degree-of-Freedom (3-DoF) movement.

3.5 Procedure and Task

The experiment employed a 3-Alternative-Forced-Choice (3AFC) paradigm using a so-called transferring task [25]. In each trial, participants were presented with three audio stimuli: two generated using a Reference method and one using a Test method (or vice versa). Crucially, to isolate room acoustics, each stimulus was convolved with a unique anechoic speech sample, preventing simple pattern matching and forcing oddball

identification based only on abstracted room acoustic features rather than direct signal-level comparison [25, 26].

Participants interacted with a floating Graphic User Interface (GUI) rendered within the virtual environment (see Fig. 3b). The interface provided dedicated playback controls ("Play 1", "Play 2", "Play 3", and "Stop"), allowing listeners to freely switch between the three stimuli at any time. To register their response, participants first identified the deviant audio sample by selecting the corresponding number (1, 2, or 3) under the prompt asking which sample featured a different room acoustic environment. Subsequently, they rated the magnitude of the perceived difference on a continuous slider anchored from "Slightly different" (0) to "Completely different" (1), before clicking a final submit button to log the trial. This subjective rating was included under the hypothesis that some acoustic deviations might be so detectable that a secondary metric would be required to quantify the perceptual penalty.

3.6 Bayesian Statistical Analysis

A Bayesian statistical framework was adopted to analyze both the detection test and the subjective rating data [27]. Detection probability in the 3AFC task was estimated using Bayesian Logistic Regression. For the trial i, the binary response Y_i (1 for correct detection, 0 for incorrect) uses a modified link function to account for the $\gamma = 1/3$ guessing floor, such that $Y_i \sim \text{Bernoulli}(p_i)$, where $p_i = \frac{1}{3} + \frac{2}{3} \cdot \frac{1}{1+e^{-\eta_i}}$. The linear predictor η_i (latent discriminability) is defined as $\eta_i = \alpha_{\text{method}}[i] + \beta_{\text{condition}}[i] + u_{\text{subject}}[i]$. A cell-means parameterization (no global intercept) was used, making α_{method} the absolute log-odds of discriminability per method. This constrains $p_i \in [1/3, 1]$, where positive α indicates easier detection.

The perceived difference magnitudes of the ratings task were modeled using a Bayesian Zero-One Inflated Beta (ZOIB) Regression for normalized ratings $R_i \in [0,1]$. ZOIB mixes discrete endpoint processes with a continuous Beta distribution. The continuous mean μ_i (intensity of perceived difference) is modeled as $R_i \sim \text{ZOIB}(\mu_i, \phi, \alpha, \nu)$ and uses a standard logit link $\mu_i = \frac{1}{1+e^{-\zeta_i}}$, with the linear predictor ζ_i defined as $\zeta_i = \gamma_{\text{method}}[i] + \delta_{\text{condition}}[i] + v_{\text{subject}}[i]$. Using a cell-means parameterization, γ_{method} represents the absolute logit-scale mean rating per method, while v_{subject} captures listener variability. Positive values of γ indicate a shift towards $\mu \to 1$ (higher perceived difference).

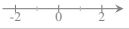
Both models were implemented using `PyMC` [28] with No-U-Turn Sampler (4 chains, 1k warm-up, 4k draws). Weakly informative priors were used: $\mathcal{N}(0, 1.5)$ for population-level fixed effects and Exponential(1.0) for participant random intercept standard deviations. Convergence was strictly confirmed ($\hat{R} \leq 1.01$, ESS > 2000), and results are reported as posterior means with 94% Highest Density Intervals (HDI).

Detection results are reported as Odds Ratios (OR), calculated from posterior log-odds, and absolute percentage-point (pp) shifts, which translate these odds into actual differences in detection probability. In practice, an OR > 1 indicates that the method was easy to identify, while an OR < 1 indicates that it was harder to detect. Perceived-difference ratings refer to the mean of the continuous Beta component (μ). To assess differences between rendering methods, pairwise contrasts were computed, and a difference was considered credible if its 94% Highest Density Interval (HDI) excluded zero. We also report the Probability of Direction (pd), which indicates the proportion of posterior draws in which the contrast shows the same direction as the mean. Furthermore, the precision parameter ($\phi \sim \text{Exp}(1)$) accounts for the rating dispersion, parameterising the Beta likelihood as $\alpha = \mu\phi$ and $\beta = (1-\mu)\phi$, with a higher ϕ indicating greater listener consensus. Although few comparisons reached strict HDI credibility in later phases, strong directional trends ($pd > 0.90$) provide clear insight into perceptual tendencies.

4 Results

Table 1 summarizes the results of the statistical analysis. Against the measured RIR (Phase 1), ISM was the easiest to detect (OR $= 5.53$) with the highest perceived difference ($\mu = 0.52$). ISM was credibly more detectable than all others, being 14.5 times more likely to be identified as artificial than the GA simulation. In contrast, the GA and Hybrid models best approximated the measured response (OR $= 0.38$ and 0.54, respectively). Their difference in detection was 4.5 pp ($pd = 0.77$), and in ratings 5.4 pp ($pd = 0.94$). Listeners' consensus peaked here ($\phi = 3.75$), suggesting high consistency when a physical measurement was present.

Table 1: Combined summary of Bayesian regression estimates for the Detection Task (logistic) and Difference Rating Task (Zero-One Inflated Beta). The table reports the posterior Mean, Standard Deviation (SD), and 94% Highest Density Interval (HDI). For the Detection Task, the Odds Ratio (OR) quantifies detectability relative to the reference. Forest plots visualize the 94% HDI ranges.

	Detection Task				Difference Rating Task		
Parameter	**Mean (SD)**	**94% HDI**	**OR**	**Plot**	**Mean (SD)**	**94% HDI**	**Plot**
PHASE 1 (Ref: Measured)							
GA	-0.96 (0.64)	[−2.13, 0.26]	0.38		-0.37 (0.57)	[−1.47, 0.67]	
Hybrid	-0.62 (0.63)	[−1.82, 0.55]	0.54		-0.14 (0.57)	[−1.25, 0.89]	
SDN	1.16 (0.67)	[−0.08, 2.43]	3.19		0.13 (0.57)	[−0.95, 1.21]	
ISM	1.71 (0.72)	[0.39, 3.08]	5.53		0.52 (0.57)	[−0.55, 1.59]	
DFDN	0.43 (0.64)	[−0.75, 1.65]	1.54		0.15 (0.57)	[−0.96, 1.19]	
Dry	0.68 (0.60)	[−0.48, 1.76]	1.97		0.11 (0.57)	[−0.97, 1.17]	
Wet	1.04 (0.60)	[−0.12, 2.14]	2.83		0.14 (0.57)	[−0.89, 1.24]	
PHASE 2 (Ref: GA)							
Hybrid	-0.85 (0.71)	[−2.21, 0.46]	0.43		-0.37 (0.62)	[−1.55, 0.79]	
SDN	0.59 (0.67)	[−0.71, 1.81]	1.80		-0.16 (0.62)	[−1.32, 1.02]	
ISM	-0.16 (0.68)	[−1.46, 1.10]	0.85		-0.27 (0.62)	[−1.43, 0.92]	
DFDN	0.24 (0.67)	[−0.97, 1.55]	1.27		0.01 (0.62)	[−1.19, 1.15]	
Dry	-0.30 (0.65)	[−1.49, 0.95]	0.74		-0.43 (0.62)	[−1.63, 0.71]	
Wet	0.13 (0.65)	[−1.09, 1.35]	1.14		-0.31 (0.62)	[−1.45, 0.86]	
PHASE 3 (Ref: Hybrid)							
SDN	0.11 (0.74)	[−1.29, 1.48]	1.12		-0.01 (0.69)	[−1.26, 1.34]	
ISM	0.44 (0.73)	[−1.00, 1.76]	1.55		-0.29 (0.69)	[−1.63, 0.97]	
DFDN	0.28 (0.73)	[−1.09, 1.67]	1.32		-0.06 (0.69)	[−1.38, 1.22]	
Dry	-0.46 (0.72)	[−1.84, 0.85]	0.63		-0.61 (0.69)	[−1.93, 0.67]	
Wet	1.28 (0.72)	[−0.10, 2.59]	3.60		0.22 (0.69)	[−1.00, 1.58]	

With GA as a reference (Phase 2), detection became more difficult. The Hybrid model remained the closest match (OR $= 0.43$). SDN (OR $= 1.80$) was 4.2 times more likely to be detected than the Hybrid model, a highly certain contrast ($pd = 0.999$). In ratings, most methods showed negative directional trends, meaning that they were generally rated as having a lower perceived difference from the GA reference. The only credibly non-zero rating contrast was DFDN versus Hybrid, though the absolute shift was very slight (0.09). Removing physical reality as a reference slightly decreased participants' precision in intensity scaling ($\phi = 3.44$).

Against the Hybrid reference (Phase 3), no credible differences emerged across methods (ISM, SDN, DFDN) in either task (absolute differences < 5 pp). However, the acoustic condition showed a greater effect than the other phases, with the reverberant "Wet" condition being 3.6 times more likely to be detected as a deviation than the treated "Dry" condition (OR $= 3.60$, $pd = 0.99$). Precision was reduced in this part of the test ($\phi = 2.87$), suggesting a weaker consensus among participants.

A Pearson correlation between detection and ratings revealed a positive relationship across all 12 conditions ($r = 0.81$, 94% HDI $[0.47, 0.94]$), driven primarily by Phase 1 variance. In Phases 2 and 3, detection probabilities clustered centrally ($0.55 - 0.75$) with ratings ≤ 0.50. Thus, even when detected, the subjective perceptual penalty of these simplified methods remains minimal.

5 Discussion

We have identified that simple reverberation models offer two primary integration pathways in VR-based architectural design: as methods to rapidly auralize unbuilt spaces during iterative early design stages, or as more efficient real-time renderers tuned to reliable acoustic references (e.g., measurements or GA simulations) for 3- or 6-DoF auralization. As demonstrated prior to the perceptual evaluation, feeding identical tabulated material absorption values into geometrically informed artificial reverberators yielded initial decay profiles that strongly deviated from the industry-standard

GA simulation. This confirms that current implementations of lightweight algorithms cannot yet substitute the predictive validity of full-scale GA simulations for unbuilt spaces. However, when their decay parameters are explicitly tuned to match a reliable reference, these efficient implementations can bypass this limitation and serve as highly effective rendering proxies. While this makes them promising candidates for 6-DoF spatial audio, enabling interactive VR walkthroughs of complex audiovisual scenes, this specific application requires further experimental validation beyond the static positions tested in this study.

The subsequent perceptual evaluation of these calibrated approximations revealed that geometric accuracy alone does not guarantee plausibility. In Phase 1, pure ISM was more easily detected as artificial, likely due to its sparse tail, thereby providing a perceptual quantification of this known limitation in VR. Conversely, the success of the Hybrid model highlights the perceptual importance of the late reverberant field, aligning with previous findings [11, 29, 30]. Directional trends for the SDN and DFDN models further corroborate this. Because a dense, reverberant tail largely drives acoustic convincingness, bypassing full calculation of early reflections can be perceptually forgiving in this context. While optimizing reverberant decay offers a scalable solution for immersive design, accurate early reflections remain critical for self-localization. Although not tested in this study, geometrically informed models that capture these early cues could offer a vital, scalable solution for interactive spatial design.

Methodologically, the experiment highlighted distinct behavioral differences between tasks. Detection was easier to infer than rating difference magnitudes, which exhibited high variance. This variance likely stems from the inherent ambiguity of rating scales without defined anchors [31]. This ambiguity mirrors early-stage design, where undefined materials and renderings without textures require users to evaluate unbuilt spaces based on internal expectations rather than direct A/B comparisons. Because detection rates rarely reached a perceptual ceiling (particularly in Phases 2 and 3), the anticipated need for this secondary magnitude rating proved less critical than initially suspected. Moreover, participants averaged 6.9 plays per trial (SD = 3.56), suggesting that perceptual penalties will be even less pronounced in ecological VR scenarios where attention is divided across multisensory architectural tasks. Consequently, computationally efficient models can be sufficiently convincing for interactive evaluation when deployed in isolation, provided they are implemented carefully to accurately reflect the global acoustic characteristics of the target space, such as reverberation time and spectral balance.

Currently, practical limitations hinder seamless implementation. Integrating these algorithms into a unified VR workflow required bridging disparate environments (Python, C#, Unity, Reaper) and combining existing repositories, VST plugins, and custom code based on theoretical frameworks. Furthermore, several methods required custom adaptations to support Ambisonics for head-tracked binaural listening. The lack of standardized, open-source, and VR-compliant implementations remains a significant barrier to both scientific reproducibility and practical architectural integration.

Finally, this framework supports an ecologically valid approach to evaluating spatial design in VR. By decoupling heavy acoustic prediction from real-time rendering, lightweight models can be driven natively on standalone VR headsets. However, generalizability is currently constrained by the use of anechoic speech from a single position and the necessary use of first-order Ambisonics, which decreases spatial discriminability. While scalability in fully interactive multi-source environments requires further investigation, this study confirms that artificial reverberation remains a convincing proxy for immersive design.

6 Summary

The experimental assessment showed directional trends between the performance of efficient reverberation methods, high-fidelity simulations, and reality; however, it does not confirm credible perceptual comparability at this sample size. Although models lacking sufficient temporal density were consistently detected, approaches incorporating dense late reverberation proved to be more robust. As demonstrated prior to calibration, these lightweight models cannot function as predictive tools, yet, provided they are tuned to known target decay characteristics, they offer highly convincing rendering proxies for immersive auralization.

The current study used 3-DoF rendering with a single speech source and a static listener position, so scalability in fully interactive 6-DoF environments requires further investigation. Future research will also

attempt to replace manual calibration with a differentiable rendering framework that optimizes material and algorithmic parameters against reference EDCs. To enhance ecological validity, subsequent experiments will shift from perceptual discrimination to complex architectural tasks using audiovisual stimuli, revealing how auralization influences spatial decision-making and practical design utility.

Acknowledgments

This work was funded by the Swiss National Science Foundation (grant no. 10000803).

References

[1] Zhou, K., "Innovative Applications of XR in Architectural Sustainability: Current Trends and Future Prospects," *Academic Journal of Science and Technology*, 10(3), pp. 69–73, 2024, doi:10.54097/nfcfab89.

[2] Vorländer, M., *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, RWTHedition, Springer International Publishing, Cham, 2020, doi:10.1007/978-3-030-51202-6.

[3] Savioja, L., "Simulation-Based Auralization of Room Acoustics," *Acoustics Today*, 16(4), pp. 48–56, 2020, doi:10.1121/AT.2020.16.4.48.

[4] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, 65(4), pp. 943–950, 1979, ISSN 0001-4966, doi:10.1121/1.382599.

[5] Borish, J., "Extension of the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, 75(6), pp. 1827–1836, 1984, doi:10.1121/1.390983.

[6] Krokstad, A., Strom, S., and Sørsdal, S., "Calculating the acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, 8(1), pp. 118–125, 1968, doi:10.1016/0022-460X(68)90198-3.

[7] Savioja, L. and Svensson, U. P., "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, 138(2), pp. 708–730, 2015, doi:10.1121/1.4926438.

[8] Benoit, A., Politis, A., Schlecht, S., and Välimäki, V., "Directional Feedback Delay Network," *Journal of the Audio Engineering Society*, 67(10), pp. 752–762, 2019, doi:10.17743/jaes.2019.0026.

[9] De Sena, E., Haciihabiboglu, H., Cvetkovic, Z., and Smith, J. O., "Efficient Synthesis of Room Acoustics via Scattering Delay Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9), pp. 1478–1492, 2015, doi:10.1109/TASLP.2015.2438547.

[10] Wendt, T., Van De Par, S., and Ewert, S., "A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation," *Journal of the Audio Engineering Society*, 62(11), pp. 748–766, 2014, ISSN 15494950, doi:10.17743/jaes.2014.0042.

[11] Martin, V., Engel, I., and Picinali, L., "Effects of Geometrical Acoustics Model Simplification on Binaural Reverberation," *IEEE Transactions on Audio, Speech and Language Processing*, 33, pp. 3480–3493, 2025, doi:10.1109/TASLPRO.2025.3597463.

[12] Brinkmann, F., Aspöck, L., Ackermann, D., Lepa, S., Vorländer, M., and Weinzierl, S., "A round robin on room acoustical simulation and auralization," *The Journal of the Acoustical Society of America*, 145(4), pp. 2746–2760, 2019, doi:10.1121/1.5096178.

[13] May, L., Farzaneh, N., Das, O., and Abel, J. S., "Comparison of Impulse Response Generation Methods for a Simple Shoebox-Shaped Room," *Acoustics*, 7(3), 2025, doi:10.3390/acoustics7030056, 56.

[14] Brinkmann, F., Lindau, A., and Weinzierl, S., "On the authenticity of individual dynamic binaural synthesis," *The Journal of the Acoustical Society of America*, 142(4), pp. 1784–1795, 2017, doi:10.1121/1.5005606.

[15] Defays, A., Safin, S., Billon, A., Decaestecker, C., Warzée, N., Leclercq, P., and Nyssen, A.-S., "Bimodal Interaction: The Role of Visual Information in Performing Acoustic Assessment in Architecture," *The Ergonomics Open Journal*, 7(1), pp. 13–20, 2015, doi:10.2174/1875934301407010013.

[16] Christensen, C. L., "Odeon, a design tool for auditorium acoustics, noise control and loudspeaker systems," in *Proceedings of Reproduced Sound 17*, pp. 137–144, 2001.

[17] Scheibler, R., Bezzam, E., and Dokmanić, I., "Pyroomacoustics: A Python package for audio room simulations and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 351–355, 2018, doi:10.1109/ICASSP.2018.8461310.

[18] Götz, G., Falcón Pérez, R., Schlecht, S. J., and Pulkki, V., "Neural network for multi-exponential sound energy decay analysis," *The Journal of the Acoustical Society of America*, 152(2), pp. 942–953, 2022, doi:10.1121/10.0013416.

[19] Dal Santo, G., Alary, B., Prawda, K., Schlecht, S. J., and Välimäki, V., "RIR2FDN: An Improved Room Impulse Response Analysis and Synthesis," in *27th International Conference on Digital Audio Effects (DAFx24)*, University of Surrey, Guildford, United Kingdom, 2024.

[20] Abel, J. S. and Huang, P., "A Simple, Robust Measure of Reverberation Echo Density," in *121st AES Convention*, Audio Engineering Society, San Francisco, CA, USA, 2006.

[21] Fontana, M., Presti, G., Fantini, D., Avanzini, F., and Reyes-Lecuona, A., "A Highly Parametrized Scattering Delay Network Implementation for Interactive Room Auralization," in *International Conference on Digital Audio Effects*, Guildford, Surrey, UK, 2024.

[22] International Organization for Standardization, "ISO 3382-1. Acoustics - Measurement of room acoustic parameters - Part 1: Performance spaces," 2009.

[23] Werner, S., Klein, F., Mayenfels, T., and Brandenburg, K., "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, Lisbon, Portugal, 2016, doi:10.1109/QoMEX.2016.7498973.

[24] Meyer-Kahlen, N., Anja, H., Schlecht, S., and Lokki, T., "Dataset of Audiovisual Speech for AR Telepresence Studies (Speech Recordings)," 2025, doi:10.5281/zenodo.15034732.

[25] McKenzie, T., Meyer-Kahlen, N., and Schlecht, S. J., "On the role of speech similarity in the detection of room acoustic differences," *The Journal of the Acoustical Society of America*, 159(2), pp. 1373–1384, 2026, doi:10.1121/10.0042354.

[26] McKenzie, T., McCormack, L., and Schlecht, S., "The role of source signal similarity in distinguishing between different positions in a room," *Journal of the Audio Engineering Society*, 2023, doi:10.5281/ZENODO.5720724.

[27] Van De Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., and Yau, C., "Bayesian statistics and modelling," *Nature Reviews Methods Primers*, 1(1), 2021, doi:10.1038/s43586-020-00001-2.

[28] Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T., and Zinkov, R., "PyMC: a modern, and comprehensive probabilistic programming framework in Python," *PeerJ Computer Science*, 9, 2023, doi:10.7717/peerj-cs.1516.

[29] Kirsch, C., Poppitz, J., Wendt, T., Van De Par, S., and Ewert, S. D., "Spatial Resolution of Late Reverberation in Virtual Acoustic Environments," *Trends in Hearing*, 25, 2021, doi:10.1177/23312165211054924.

[30] Neidhardt, A., Schneiderwind, C., and Klein, F., "Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms - Literature Review and Theoretical Framework," *Trends in Hearing*, 26, 2022, doi:10.1177/23312165221092919.

[31] Chen, Q. Z., Weld, D. S., and Zhang, A. X., "Goldilocks: Consistent Crowdsourced Scalar Annotations with Relative Uncertainty," *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), 2021, doi:10.1145/3476076.