OPEN ACCESS Freely available online

K. Y. Lee, N. Meyer-Kahlen, S. J. Schlecht, and V. Välimäki,
"Evaluating Reverberation Models for Augmented Reality,"
J. Audio Eng. Soc., vol. 73, no. 10, pp. 619–632 (2025 Oct.).
https://doi.org/10.17743/jaes.2022.0223.

Evaluating Reverberation Models for Augmented Reality

KYUNG YUN LEE,[1,*] **NILS MEYER-KAHLEN,**[1] *AES Associate Member*,
(kyung.y.lee@aalto.fi) (nils.meyer-kahlen@aalto.fi)

SEBASTIAN J. SCHLECHT,[2] **AND VESA VÄLIMÄKI,**[1] *AES Fellow*
(sebastian.schlecht@fau.de) (vesa.valimaki@aalto.fi)

[1]*Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland*
[2]*Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany*

The ultimate goal of the reverberation model for augmented reality is to create an auditory illusion, making simulated sound sources indistinguishable from real, measured ones. However, existing evaluation methods are not tailored to achieve this objective. This paper adopts the evaluation paradigm of auditory illusion tests to evaluate reverberation models under two distinct tasks: authenticity and transferring. The listening test uses the three-alternative forced-choice design, where subjects are asked to detect the speech signal processed with a model-generated room impulse response (RIR) among the signals processed with measured RIRs. For the authenticity task, the three signals contain the same speech sample, while for the transferring task, they contain different samples from different speakers. A Bayesian analysis shows that detecting model-generated RIRs is significantly more challenging in the transferring task than in the authenticity task across all models. Additionally, while the listening test results positively correlate with the selected objective metrics, the reliability and generalizability of these correlations for predicting listening test outcomes remain uncertain. The proposed evaluation framework for reverberation models can serve as a precursory analysis for developing dynamic, binaural rendering for augmented reality applications.

0 INTRODUCTION

Models of room impulse responses (RIRs) have widespread utilities, ranging from music production [1] and auralization in room acoustic design [2, 3] to augmented reality (AR) and virtual reality [4, 5]. In general, a reverberation model aims to encode the room acoustics of a given space using a limited set of parameters, which can then be used to generate responses that, in turn, are used to process source signals. Although numerous models of varying complexities have been introduced, their perceptual fidelity is often assessed without much attention to the application [6–9]. This paper uses a well-defined, reproducible paradigm to evaluate the perceptual fidelity of reverberation models for AR applications.

Reverberation models can range from a simple decaying noise generator with reverberation time (RT60) as a parameter to a deep learning model with tens of thousands of coefficients. Regardless of complexity, any reverberation model can be thought of as comprising an analyzer and a generator. The analyzer extracts a set of parameters from the target RIR, and the generator resynthesizes the RIR from the parameters, see Fig. 1. Thus, the goal of the model is to capture the target RIR as closely as possible, typically aiming to achieve this with the smallest possible set of parameters while making necessary compromises. Depending on priorities, the model can be geared toward computational efficiency, flexibility, or interpretability.

Thus far, model evaluation is often conducted by letting subjects directly compare the degree of similarity between the model-generated and target RIRs [6–9], for example, in a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)–like test [10]. Therein, different model outputs are directly compared to each other and the target signal. All stimuli are rendered with the same source signal. To reach the maximal score, the model must generate RIRs that are indistinguishable from the target under all conditions, as such a direct comparison makes even minor differences audible. Since, in MUSHRA-like tests, several models or variants thereof are compared to each other directly, the selection of the included models, the anchor, and

*To whom correspondence should be addressed, email: kyung.y.lee@aalto.fi.

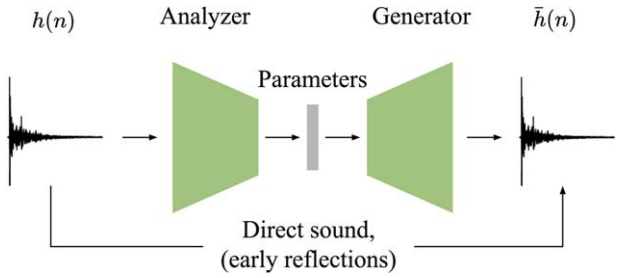


Fig. 1. A general reverberation model consists of an analyzer and generator. It is common that reverberation models do not generate the direct sound path, while some do generate the early reflections.

the type of source signal have a large influence on the results of the test. More importantly, a situation involving a direct comparison does not occur in the real world, making it an unnecessarily difficult evaluation for many practical applications. Nevertheless, it remains valuable for applications that require high-fidelity results.

Since in AR systems, real and virtual sound sources rarely emit the same source signal, a type of indirect comparison is more relevant there. For example, in a multi-speaker scenario with both real and virtual people, no two signals contain the same person uttering the same sentence. In the context of AR, evaluation methods relying on different types of indirect comparison involving real and virtual signals have been referred to as plausibility [11] and transfer-plausibility [12], whereas those incorporating direct comparison indicate authenticity [13]. Recent evaluations for AR applications adopt real/virtual tests to determine if an auditory illusion is created (i.e., if virtual sources are believed to be real) using AR equipment. These tests have proven effective in comparing room acoustic rendering conditions [14]. However, as mentioned earlier, they require dedicated AR equipment and high-quality binaural rendering to start with.

Instead, a virtual-only test can be performed for appropriate situations, such as evaluating nonbinaural signals. For instance, reverberation models from acoustic simulation algorithms have been evaluated using this method [15]. A virtual-only test can also serve as a precursory evaluation method to the real/virtual test. The difference is that in the virtual-only test, only headphones are used instead of AR equipment, and the reference signal is a measured real-world signal played over headphones instead of a loudspeaker. An example of a virtual-only test is the classical direct-comparison MUSHRA-like test. Meanwhile, a virtual-only test involving indirect comparison is referred to as a "transferring task" [16, 17]. McKenzie et al. [16] have shown that differences between positions in a room are harder to detect in such a transferring task compared to a direct comparison task. Additionally, Kuusinen and Lokki [18] have demonstrated that comparing concert hall acoustics is much more difficult when different excitation signals are used for the conditions than when the same excitation signal is used.

Despite the development of practical paradigms, there has been limited work specifically evaluating reverberation models under these paradigms. Fantini et al. [19] investigated the perception of reverberation models with one type of late reverberation simplification under a similar paradigm, which they refer to as coimmersion, showing that subjects had difficulties detecting the simplified model. Moreover, Schneiderwind et al. [20] experimented with the degree of late reverberation modification that affects externalization and audiovisual plausibility in a real/virtual test.

This work is designed to observe the gap in the perceptual fidelity of static monoaural reverberation models between the authenticity task and transferring task. The listening test follows a three-alternative forced-choice (3-AFC) design, where participants perform a detection task to evaluate reverberation models under uncertainties introduced by varying levels of source similarity. Here, the source signal refers to speech signals, which are processed with model-generated RIRs. The two tasks differ in their source type conditions: In the authenticity task, RIRs are compared using the same speaker, whereas the transferring task involves different speakers, increasing detection difficulty. If model differences emerge in the transferring task, they can be considered more relevant to practical applications than if they were just found through direct comparison. Evaluating static monaural RIRs under this task serves as the first step in achieving auditory illusion for dynamic reverberation models in AR applications.

Here, the listening test results are analyzed with a Bayesian model that estimates the probability of correctly detecting the model-generated RIR. The posterior distributions reveal how various factors influence detection rates and how detection rates depend on the two source signal types. In addition, an objective evaluation is performed using well-known metrics, and the correlation between the objective and perceptual evaluations is demonstrated.

Six reverberation models of varying complexities are selected to address the following research questions (RQs):

- RQ1: Is it easier to detect model-generated RIRs in the authenticity than in the transferring task?
- RQ2: Do certain rooms make detecting the model easier?
- RQ3: How important is it to model the early reflections accurately?
- RQ4: Do perceptual and objective evaluations correlate?

The paper is structured as follows. SEC. 1 introduces the listening test design, reverberation models, and dataset that were used. Then, the analysis and results are presented in SEC. 2 and are further discussed in SEC. 3. SEC. 4 concludes. The code for experiments and models is provided as a baseline for future research: https://github.com/kyungyunlee/rir-model-eval.

1 PROPOSED METHOD

The proposed evaluation framework is specifically designed to evaluate models independently, without facili-

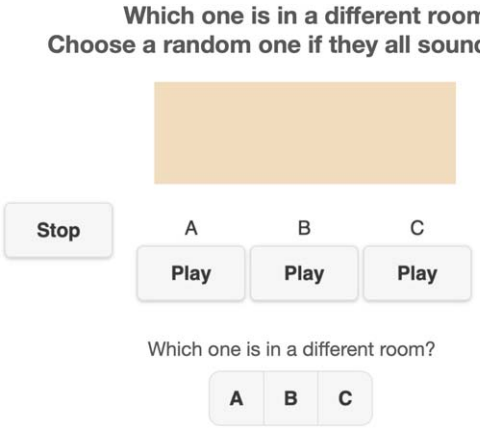


Fig. 2. Screenshot of the listening test interface. The webMUSHRA template was modified to create a 3-AFC design. No speech waveforms were displayed.

tating comparisons between them. This design allows the framework to be applied flexibly, either to evaluate a single model or to independently assess multiple models within the same session. As an example, six models are evaluated, detailed in SEC. 1.3.

1.1 Listening Test Design

A 3-AFC design detection task is proposed (Fig. 2). For every trial, subjects are presented with three reverberated speech signals, consisting of two speech signals convolved with the target RIR and one speech signal with the model-generated RIR from one of the models. The task is to detect the one containing the model-generated RIR, answering "which one is in a different room?" In case of difficulty in detection, subjects are asked to make a random choice. Since a single model is tested during each trial, there is no immediate comparison between models.

To investigate RQ1, the experiment consists of a random sequence of authenticity and transferring tasks, in which the source signal type differs. Here, the source type refers to the selection of the anechoic speech signal, which will be convolved with RIRs. For the authenticity task, three RIRs are convolved with the same source signal—from the same speaker uttering the same sentence. On the other hand, for the transferring task, all three RIRs are convolved with different signals—from different speakers uttering different sentences [16].

Model performance is summarized using the detection rate. A high detection rate indicates an undesirable scenario in which identifying the model-generated RIR was easy. The indistinguishability of model-generated RIRs from the measured RIR can be defined using the guessing rate. In the 3-AFC task, a detection rate of $\frac{1}{3}$ is the guessing rate, indicating that subjects are unable to distinguish between the model-generated and target RIR, leading them to make random choices. In other words, this value is the limit of transfer-plausibility and can be interpreted as the optimal score that a successful model can achieve.

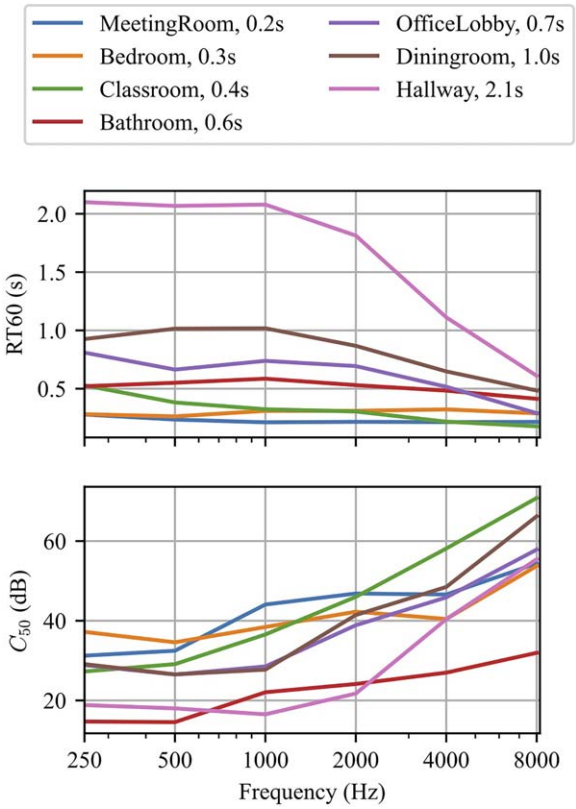


Fig. 3. RIR parameters (RT60 and clarity C_{50}) across frequency bands are shown for RIRs used in the dataset. Corresponding RT60 values are included in the legend. All RIRs are normalized to have the same direct sound peak level.

With both the same and different speakers, each model is evaluated with seven different RIRs discussed in SEC. 1.2. With six models, this experiment consists of 84 trials in total, half of which tests with the same speakers, whereas the other half uses different speakers. For every subject, a unique version of the test was rendered with a new set of randomized speakers to reduce any systematic influence of the specific speaker. All tests were conducted in the listening booth at the Aalto Acoustics Lab with a Sennheiser HD 650 headphone. The playback level was set to a comfortable listening level, which was measured to be $L_{A,\,eq} \approx 64$ dB at the ear simulator of a KEMAR head and torso simulator. The level was fixed and unchanged for all subjects. The testing software was a modification of the webMUSHRA interface.

1.2 Test Dataset

Seven RIRs are selected from the MIT IR Survey dataset [21] with varying RT60 profiles across frequency (Fig. 3). This dataset was chosen for its diverse range of measurement environments from domestic to outdoors. The original sampling rate of the dataset is 32 kHz, but signals are resampled to 48 kHz for this paper. The anechoic speech signals are recordings made at the Aalto Acoustics Lab with the Harvard speech dataset [22]. The dataset contains 21 speakers uttering 20 sentences each.

Table 1. List of reverberation models under evaluation. Brief descriptions of the analyzer and generator structure and the types of parameters are presented.

Model	Analyzer	Parameters	Generator	Early Reflections
`RT2RIR-LateOnly`	DecayFitNet	RT60s, amplitudes	Decaying noise generator	Copied
`RT2RIR`	DecayFitNet	RT60s, amplitudes	Decaying noise generator	Generated
`EchoDensity`	Echo density profile, DecayFitNet	AED, RT60s, amplitudes	Inverse echo density, Decaying noise generator	Generated
`FDN`	DecayFitNet	RT60s, amplitudes	FDN	Generated
`DarkVelvetNoise`	Least squares optimizer	Filter probabilities	Weighted dictionary filters	Copied
`FiNS`	Neural network	Latent acoustic features	Neural network	Generated

1.3 Tested Reverberation Models

All models described in this paper are implemented to work with a single-channel static RIR at a sampling rate of 48 kHz. As the RIRs from the aforementioned dataset are bandlimited to 16 kHz, models are designed to synthesize only up to 16 kHz. For all models, the direct sound path is not generated but is copied from the target RIR. Table 1 shows the overview of all models in terms of how each model is designed. Fig. 4 presents the short-time Fourier transforms (STFTs) of model-generated RIRs for the "Classroom" RIR.

All models, except `Filtered Noise Shaping (FiNS)`, appear to have synthesized late reverberations with similar decay rates as that of the measurement across frequency bands. In particular, `RT2RIR-LateOnly`, `RT2RIR` and `EchoDensity` show very similar late reverberation patterns because they use the same late reverberation model. However, particularly visible from that of `EchoDensity` and `Feedback Delay Network (FDN)`, early reflections do not appear to have been generated as accurately as the late reverberation, which is expected as early reflections are known to be more difficult to model due to their nonstatistical characteristics.

1.3.1 Exponentially Decaying Noise Model

The simplest reverberation model generates an exponentially decaying noise with a decay rate corresponding to the RT60 in octave bands. Apart from the direct sound path, the entire RIR is generated in such a way and is referred to as `RT2RIR`. In addition, `RT2RIR-LateOnly` is a model that generates only the late reverberation from noise, while copying the early reflections from the target RIR. `RT2RIR-LateOnly` and `RT2RIR` models were both included in the study to investigate the effect of accurate early reflection modeling, addressing RQ3.

For both `RT2RIR` and `RT2RIR-LateOnly`, the analyzer computes RT60 and amplitude per octave band from the target RIR using the DecayFitNet estimator [23]. The analysis is performed over six octave bands selected through the pilot listening test: 250; 500; 1,000; 2,000; 4,000; and 8,000 Hz. The estimated RT60 values and the amplitudes are then used by the generator to set the decay rate and the initial energy level of the Gaussian noise per frequency band b [24],

$$\bar{h}_{\text{Late}}(b,t) = A_b \exp\Big(-\frac{\delta_b t}{f_s}\Big) \tag{1}$$

$$\delta_b = -\frac{\ln(10^{-6})}{\mathrm{T}_b}, \tag{2}$$

where A_b, δ_b, and T_b refers to the initial amplitude, decay rate, and RT60 value of the band b, respectively; f_s denotes the sampling rate; and t is time in samples. Then, the output of the generator is the sum of decaying noises from all bands [24],

$$\bar{h}_{\text{Late}}(t) = \sum_{b=1}^{B} \bar{h}_{\text{Late}}(b,t), \tag{3}$$

where B is the number of frequency bands.

In case of `RT2RIR-LateOnly`, the early reflection segment from the target RIR is concatenated with the synthesized decaying noise at the crossover time, which was arbitrarily determined by the average RT60 value: If RT60 is less than 0.5 s, the crossover time was set to 50 ms; otherwise, it was set to 100 ms.

In this work, `RT2RIR-LateOnly` and `RT2RIR` are considered baselines due to their simplicity. Among models that only generate the late reverberation, `RT2RIR-LateOnly` is the simplest. For models that generate both the early reflections and the late reverberations, `RT2RIR` is considered to be the most primitive model.

1.3.2 Echo Density Model

The echo density model represents RIRs by the number and the distribution of reflections over time in a room [25]. Specifically, the normalized echo density (NED) indicates the fraction of RIR samples that lie outside the Gaussian distribution within a time window. NED initially starts near 0, but as time progresses, it converges to 1 as the reflections in a room reach a diffuse state, resembling a Gaussian statistic. The first moment in time when NED reaches a value of 1 is used as the mixing time. Another measure, absolute echo density (AED), computes the number of echoes arriving at the receiver per second. Using AED, echo sequences of the original RIR can be resynthesized by sampling the wait time between echoes from an exponential distribution and the echo gain from a Gaussian distribution scaled with the previous echo gain [26].

This paper follows the echo density model of [26] and [27], where the early reflections are modeled with the echo density measures and the late reverberation is modeled with decaying Gaussian noise. The analyzer of the `EchoDensity` model computes the AED and the DecayFitNet pa-

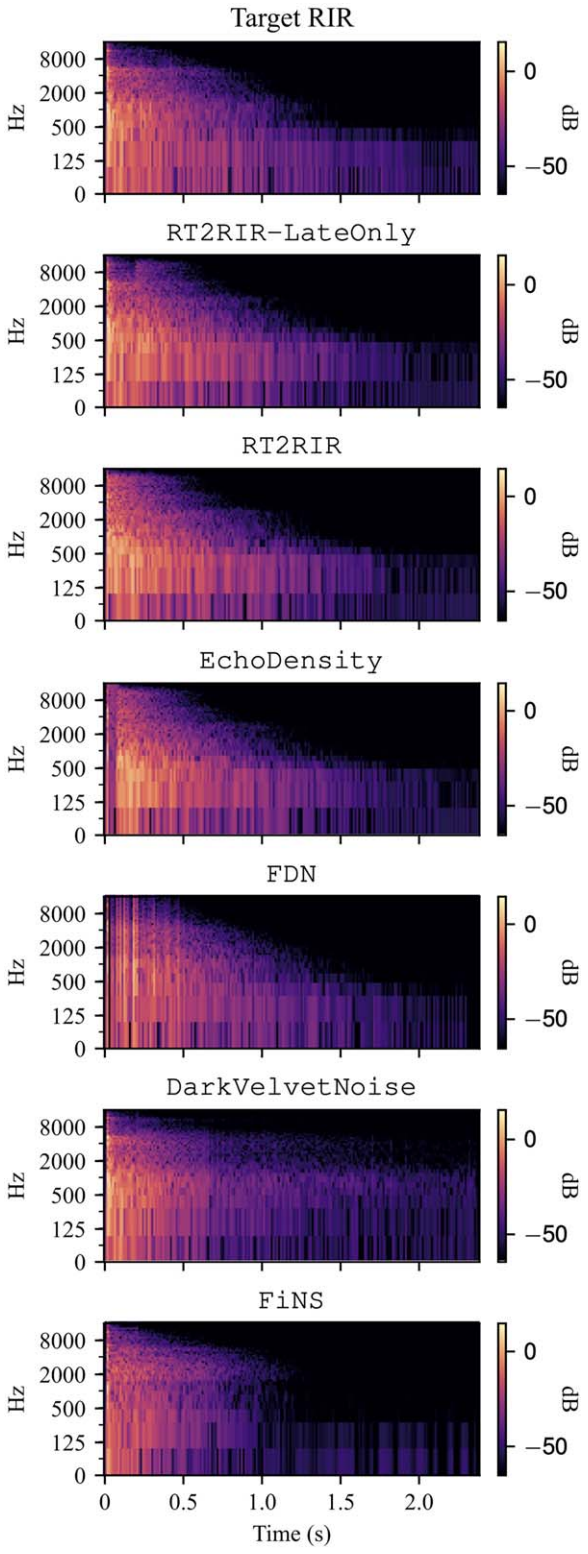


Fig. 4. STFT of model-generated RIRs for "Classroom" RIR. The target RIR is displayed at the top. The "Classroom" RIR has an average RT60 value of 0.3 s.

rameters. As in [26] and [27], the generator synthesizes the early reflections as a sequence of echoes from AED, and the late reverberation is synthesized in the same way as in `RT2RIR` with the DecayFitNet parameters.

1.3.3 Feedback Delay Network

An FDN is a recursive system that is commonly used for generating artificial reverberation [28–31]. It consists of parallel delay lines that are interconnected with a feedback matrix that controls the gains from each connection. There are many design choices that can be made in constructing the FDN. By setting appropriate parameters of the FDN, it is possible to generate reverberation with desired characteristics [29, 32, 30].

This work evaluates the `FDN` model, which uses FDN as the generator. As in the aforementioned models, the analyzer estimates RT60 with the DecayFitNet [23] to obtain the RT60 and amplitude values. These values are then used to set the parameters of the FDN with 16 delay lines and a random orthogonal feedback matrix. This model generates both the early reflections and the late reverberation. The implementation is a Python translation of Schlecht's MATLAB FDN Toolbox [32].

1.3.4 Filtered Noise Shaping Model

The FiNS model [7] is a neural network model based on recent work for blind estimation of RIRs. It is motivated by how the late reverberation can be simplified as a sum of decaying noise sequences. Thus, the model predicts the time-domain decay envelope, instead of predicting sample-by-sample values of the reverberation. Since the direct sound path and early reflections do not exhibit any statistical characteristics, these are generated per sample.

Although the original FiNS model is designed for a blind estimation task [7], it is modified so that the model resembles an autoencoder. The modified model is referred to as `FiNS` model. Now, the input to `FiNS` is 1 s of the target RIR, instead of 2.73 s of reverberant speech. To accommodate the reduced input size, the number of layers is reduced from 14 to 13, and all channel sizes are reduced by half. The multiresolution STFT loss [33], was utilized as in the original model. To train the `FiNS` model, the FRL [34] dataset, a subset of the Arni [35], and the MIT IR Survey dataset (excluding the RIRs used for the evaluation) [21] were used. The model is trained for 700 epochs with a batch size of 16. The weights used for this paper are provided in the accompanied code.

1.3.5 Dark Velvet Noise Reverberator

Velvet noise is a randomized sparse sequence of unit impulses parametrized with impulse density [36]. Velvet noise with an average impulse density of 1,500 impulses per second is shown to be the smoothest, even smoother than the Gaussian white noise [36]. Extending original velvet noise, dark velvet noise is characterized by possessing a low-pass spectrum, which is achieved by replacing unit impulses with pulses of random width [37].

An implementation was used in which the dark velvet noise is used to generate late reverberation close to that of the target RIR [38]. This method utilizes a predetermined set of filters that are weighted by probabilities to synthesize reverberation. The probabilities can be interpreted as the spectral evolution of the reverberation model. During the

analysis stage, filter probabilities are computed by running non-negative least-squares optimization on the target RIR [38]. Then, in the generator stage, the model outputs the dark velvet-noise sequence, assigns a "dictionary filter" for each pulse in the sequence based on the computed filter probabilities, and applies the weighted filter [38]. To be consistent with the `RT2RIR-LateOnly` model that does not generate the early reflections, the mixing time was set in the same way as for `RT2RIR-LateOnly`. This model is referred to as `DarkVelvetNoise`.

1.4 Bayesian Listening Test Analysis

A Bayesian approach was adopted to analyze the results of the listening test, which allows for quantifying uncertainty, particularly given the limited sample size, and improving the interpretability of how each factor, such as the source type, influences the detection rate.[1] Specifically, a Bayesian logistic regression model was implemented to estimate a binary categorical response variable, Y, which predicts whether the subject will detect the model-generated RIR correctly or not for a given trial.

To estimate Y, the model incorporates four predictors that influence the detection rate: Source Type, Reverberation Model, Target RIR, and Subject. Each predictor is a categorical variable. Source Type has two categories, same and different speakers, which are parametrized by δ_1 and δ_2, respectively. Reverberation Model has six categories, one for each of the tested reverberation models described in SEC. 1.3. The effect of each reverberation model is represented by γ. Target RIR has seven categories for the seven RIRs used in this paper, parametrized by θ. Lastly, Subject has 20 categories for the 20 subjects who participated in the listening test, with β as parameters. All the parameters, δ, γ, θ, and β, are latent random variables, inferred from the listening test data. Counting all possible combinations of the aforementioned categories leads to a total of 1,680 unique trials.

Trial i is one of the 1,680 combinations, where sourcetype[i] indicates if the trial i uses the same or different speakers, reverberationmodel[i] indicates which reverberation model is used, and so forth. Let Y_i be the 0/1 indicator of whether the subject correctly detected the model-generated RIR or not for a given trial i. Then, the model fit is

$$\begin{aligned} Y_i \mid \delta, \gamma, \theta, \beta &\sim \text{Bernoulli}(p_i) \\ \text{logit}_i &= \delta_{\text{source type}[i]} + \gamma_{\text{reverberation model}[i]} \\ &\quad + \theta_{\text{target RIR}[i]} + \beta_{\text{subject}[i]} \\ p_i &= \tfrac{1}{3} + (1 - \tfrac{1}{3})\tfrac{1}{1+e^{-\text{logit}_i}}, \end{aligned} \tag{4}$$

where p_i is the detection rate of trial i computed using the logit link function. To accommodate the 3-AFC paradigm, where the guessing rate is $\frac{1}{3}$, the standard sigmoid function is shifted to ensure that the estimated probability remains within the expected range of $[\frac{1}{3}, 1]$. This formulation aligns with the psychometric function.

This Bayesian logistic regression model provides an intuitive and probabilistically coherent framework to quantify how each parameter influences the detection rate. For example, the parameters of the Reverberation Model can be interpreted as follows: If γ_1 has a higher mean than γ_2, then `RT2RIR-LateOnly` is less capable of accurately generating the target RIR compared to `RT2RIR`. Thus, γ can be interpreted as a measure of model infidelity, where higher values indicate greater deviation from the target RIR and poorer performance, while lower values suggest better model fidelity. Similarly, the parameters of the Subject can be viewed as a measure of subject acuity, where a higher mean of β indicates that the subject is more proficient at detecting model-generated RIRs.

In summary, a lower logit value corresponds to a lower detection rate, with a logit value of 0 yielding a detection rate of $\frac{2}{3}$. A 1-unit increase in logit results in a multiplicative increase in odds by a factor of $e^1 = 2.7$. This means that an increase of 1 unit in δ makes correct detection 2.7 times more likely, while a decrease of 1 unit reduces the odds by the same factor.

To incorporate prior knowledge and prevent overfitting, prior distributions for the model parameters are defined as follows:

$$\begin{aligned} \delta_1(\text{different}) &\sim \mathcal{N}(-0.5, 0.5^2) \\ \delta_2(\text{same}) &\sim \mathcal{N}(0.5, 0.5^2) \\ \gamma_1, \gamma_2, ..., \gamma_6 &\sim \mathcal{N}(0.0, 0.5^2) \\ \theta_1, \theta_2, ..., \theta_7 &\sim \mathcal{N}(0.0, 0.2^2) \\ \beta_1, \beta_2, ..., \beta_{20} &\sim \mathcal{N}(0.0, 0.2^2), \end{aligned} \tag{5}$$

where $\mathcal{N}(\mu, \sigma^2)$ indicates the normal distribution with the mean μ and standard deviation σ. The selection of priors was based on prior belief and empirical validation through prior predictive checks and sensitivity analysis. Specifically, it is assumed that evaluating with different speakers is more challenging than with the same speaker, encoding this prior belief by setting their prior means to -0.5 and 0.5, respectively.

The logistic regression model was implemented using PyMC [41] and estimated the parameters via Markov Chain Monte Carlo sampling, employing the No-U-Turn Sampler. Four independent Markov Chain Monte Carlo chains, each consisting of 4,000 iterations, were run with 20% (800 samples) discarded as tuning iterations. To assess convergence, trace plots, R-hat values, and effective sample sizes (ESSs) were examined. The recommended thresholds of R-hat < 1.05 and ESS $> 10{,}000$ were met [40] for all parameters, ensuring reliable posterior estimates. Posterior means, standard deviations, 94% highest posterior density intervals (HPDI), and odds are reported. HPDI represents the shortest interval containing a specific percentage of the posterior probability mass for a given parameter. This is different from the frequentist confidence interval, which is based on repeated sampling because HDPI directly quantifies the uncertainty about the parameter after a single analysis.

[1] For an overview of Bayesian statistics, refer to [39, 40].

Table 2. Summary statistics of the parameter estimates for the logistic regression model. All parameter estimates converged with ESS greater than 10,000 and an R-hat equal to 1.00. The plot visualizes the 94% HPDI with the mean marked with the circle.

Parameter (Category)	Mean	SD	94% HPDI	Odds	Plot
SOURCE TYPE					−2 −1 0 1 2
δ_1 (same)	1.27	0.21	[0.88, 1.69]	3.57	
δ_2 (different)	-0.74	0.22	[-1.16, -0.34]	0.48	
REVERBERATION MODEL					
γ_1 (`RT2RIR-LateOnly`)	-0.48	0.25	[-0.95, -0.02]	0.62	
γ_2 (`RT2RIR`)	0.58	0.24	[0.13, 1.04]	1.79	
γ_3 (`EchoDensity`)	0.76	0.25	[0.31, 1.24]	2.15	
γ_4 (`FDN`)	-0.03	0.24	[-0.51, 0.41]	0.97	
γ_5 (`DarkVelvetNoise`)	-1.21	0.27	[-1.72, -0.72]	0.30	
γ_6 (`FiNS`)	0.93	0.25	[0.46, 1.39]	2.54	
TARGET RIR					
θ_1 (Bathroom)	-0.05	0.16	[-0.34, 0.25]	0.95	
θ_2 (Bedroom)	0.02	0.15	[-0.26, 0.32]	1.02	
θ_3 (Classroom)	-0.12	0.15	[-0.40, 0.17]	0.89	
θ_4 (Dining room)	-0.01	0.15	[-0.31, 0.27]	0.99	
θ_5 (Hallway)	0.13	0.16	[-0.16, 0.42]	1.14	
θ_6 (Office lobby)	0.31	0.16	[0.01, 0.60]	1.37	
θ_7 (Meeting room)	-0.20	0.15	[-0.49, 0.09]	0.82	
SUBJECT					
β_1 (Subject 1)	0.00	0.18	[-0.34, 0.32]	1.00	
β_2 (Subject 2)	0.06	0.18	[-0.27, 0.40]	1.07	
β_3 (Subject 3)	-0.24	0.17	[-0.56, 0.08]	0.79	
β_4 (Subject 4)	0.07	0.18	[-0.25, 0.41]	1.08	
β_5 (Subject 5)	-0.03	0.18	[-0.36, 0.31]	0.97	
β_6 (Subject 6)	-0.11	0.17	[-0.45, 0.20]	0.90	
β_7 (Subject 7)	0.20	0.18	[-0.13, 0.55]	1.23	
β_8 (Subject 8)	-0.03	0.17	[-0.36, 0.29]	0.97	
β_9 (Subject 9)	-0.04	0.18	[-0.38, 0.29]	0.96	
β_{10} (Subject 10)	-0.06	0.17	[-0.38, 0.28]	0.95	
β_{11} (Subject 11)	0.03	0.17	[-0.29, 0.35]	1.03	
β_{12} (Subject 12)	0.03	0.17	[-0.30, 0.36]	1.03	
β_{13} (Subject 13)	0.02	0.17	[-0.31, 0.35]	1.02	
β_{14} (Subject 14)	0.22	0.18	[-0.11, 0.57]	1.24	
β_{15} (Subject 15)	0.02	0.17	[-0.30, 0.35]	1.02	
β_{16} (Subject 16)	-0.10	0.18	[-0.43, 0.24]	0.91	
β_{17} (Subject 17)	0.01	0.18	[-0.32, 0.35]	1.01	
β_{18} (Subject 18)	-0.13	0.17	[-0.47, 0.17]	0.88	
β_{19} (Subject 19)	0.08	0.18	[-0.26, 0.42]	1.09	
β_{20} (Subject 20)	-0.05	0.18	[-0.37, 0.28]	0.95	

2 RESULT

2.1 Listening Test Results

The listening test involved 20 subjects (16 male, four female) aged 25 to 43 years, with 19 subjects having more than 1 year of experience in acoustics research. The results, summarized in Table 2, provide insights into how each predictor variable influences detection rate. The plot column visualizes the relative impact of these parameters to highlight which ones influence the detection rate more than others. A mean value close to 0 suggests that the parameter has little or no effect on the detection rate, while a negative

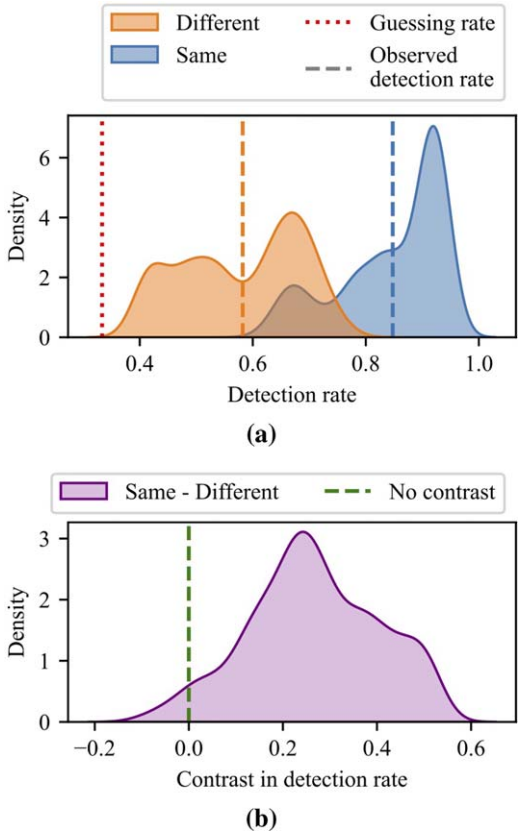


Fig. 5. (a) Posterior distributions of detection rates for the same and different speaker trials. The dashed lines represent observed detection rates of using different and same speaker, computed from the listening test results, and the dotted red line indicates the 3-AFC guessing rate. The multimodal shape of the distributions arises from variations in model fidelity. (b) Posterior probability contrast represents the difference between the distributions of detection rates of the same and different speaker trials in (a). In (b), the distribution is unimodal, as the variation now stems solely from the effect of source type, and 97% of it lies to the right of 0.0, indicating a 0.97 probability that the subject is more likely to make a correct detection in a same speaker trial.

mean indicates that detection becomes more difficult, and a positive mean suggests easier detection. If both bounds of the 94% HPDI are entirely above or below 0, the parameter has a strong effect. Among all predictors, Source Type exhibited the strongest predictive power, followed by Reverberation Model, while Target RIR and Subject had little impact.

Since the effect of source type is the primary focus of this study, the posterior distribution of detection rate was analyzed for the same and different speakers, as shown in Fig. 5(a). The results indicate that using different speakers generally leads to a lower detection rate compared to using the same speaker. The posterior probability contrast in Fig. 5(b) further quantifies this difference, showing that the probability of detection being higher when evaluated with the same speaker is 0.97. This effect is also evident in the odds (Table 2): Subjects were 3.57 times more likely to correctly detect the model-generated RIR when listening to the same speaker, whereas for different speakers, they were only 0.48 times as likely. This highlights the increased difficulty in detecting the model-generated RIR when the sources are dissimilar.

When examining different reverberation models (Fig. 6), it was also observed that detection was consistently more difficult when evaluated with different speakers, because the maximum a posterior (MAP) of the different speaker distribution is significantly lower than that of the same speaker distribution. Table 2 shows that `RT2RIR-LateOnly` and `DarkVelvetNoise` are the best-performing models, with posterior means well below 0 and both bounds of the 94% HPDI in the negative range. In contrast, `RT2RIR`, `EchoDensity`, and `FiNS` are easier to detect, with posterior means above 0. `FDN` shows no strong effect, suggesting it is neither particularly easy nor difficult to detect.

Variations in target RIRs have a minor impact, except for *OfficeLobby*, where detection tends to be slightly easier. In response to RQ2, the posterior distributions of the θ parameters variables indicate that the variations of the RIRs used for evaluation had minimal impact on detection difficulty in this experiment. No specific RIR significantly enhanced or impaired detection. Lastly, subject acuity shows a minor effect, as posterior distributions of all β parameters are centered around 0.

2.2 Objective Metrics

Models are also evaluated on objective metrics to compare with the perceptual evaluation from the listening test. Three metrics are considered: mean-squared-error (MSE), multiresolution STFT [33], and pychoacoustic loss [42]. To align with the result of the listening test, all losses are computed between the reverberated speech signals rendered with the target RIR and model-generated RIRs. As the most naive and straightforward metric, the MSE is computed as

$$y(t) = x(t) * h(t), \tag{6}$$

$$\bar{y}(t) = x(t) * \bar{h}(t), \tag{7}$$

$$\mathrm{MSE} = \frac{1}{T}\sum_{t=0}^{T-1}(y(t) - \bar{y}(t))^2, \tag{8}$$

where * is a convolution operation; x is an anechoic speech signal; h and $\bar{h}$ are target RIR and model-generated RIR, respectively; and T is the length of the rendered signal in samples.

Multiresolution STFT is an aggregated STFT loss of spectral convergence loss L_{sc} and log magnitude loss L_{mag} over M resolutions [33]

$$L_{\mathrm{sc},m}(y, \bar{y}) = \frac{\|\mathrm{STFT}_m(y) - \mathrm{STFT}_m(\bar{y})\|_{\mathrm{F}}}{\|\mathrm{STFT}_m(y)\|_{\mathrm{F}}}, \tag{9}$$

$$L_{\mathrm{mag},m}(y, \bar{y}) = \frac{1}{T}\|\log(\mathrm{STFT}_m(y)) - \log(\mathrm{STFT}_m(\bar{y}))\|_1, \tag{10}$$

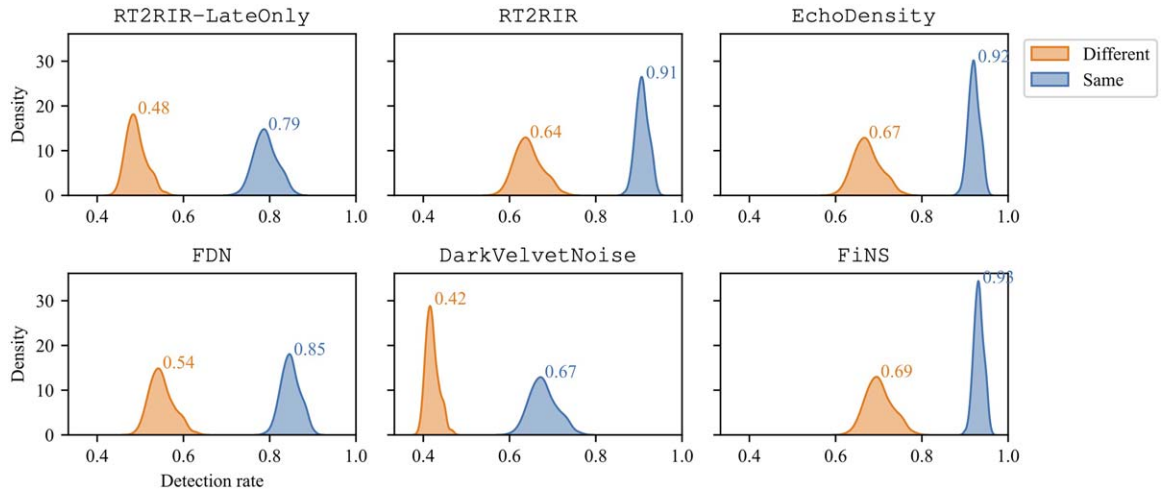


Fig. 6. Posterior distributions of detection rates for each model. The distributions for the same and different speakers are plotted separately. The MAP of each distribution is displayed adjacent to it.

$$L_{\text{stft}} = \frac{1}{M}\sum_{m=1}^{M}(L_{\text{sc},m}(y,\bar{y}) + L_{\text{mag},m}(y,\bar{y})), \tag{11}$$

where $\|\cdot\|_1$ and $\|\cdot\|_F$ refer to the L_1 and the Frobenius norm, respectively. The implementation provided by Steinmetz and Reiss [43] is used, and $M = 3$ at 512; 1,024; and 2,048-sample FFT sizes with corresponding window sizes of 256, 512, and 1,024 samples are chosen.

The psychoacoustic loss is a perceptual loss function for training deep learning models, where auditory masking is incorporated in order to minimize the spectral differences between two signals [42].[2] It applies a filter on the signal, which is a frequency-dependent normalization based on the masking threshold of human hearing per Bark band. After filtering the signal, the psychoacoustic loss computes the MSE between the given signals.

For each metric, the presented loss value is an average of 21 speech signals (one sentence from each speaker) from the speech dataset (SEC. 1.2). Fig. 7 shows the results across different rooms. In general, `RT2RIR-LateOnly` and `DarkVelvetNoise` show the lowest error on average across all metrics, which is expected as these models contain the exact early reflections of the target RIR. The errors for other models exhibit a similar trend across the rooms. In particular, noticeably high errors are shown for the RIR of "Bathroom" for these models, which may be due to the unique early reflections that are difficult to capture. However, the values obtained from these objective metrics alone are difficult to interpret with regard to the perceptual quality.

To access RQ4, adjusted R-squared (R^2_{adj}) values are computed to analyze the relationship of the same and different speaker results to the objective metrics (Fig. 8). All objective metrics are log-transformed to be in a decibel scale before performing the linear regression fitting. Lines represent the regression fits, and the shaded areas indicate the 95% confidence interval for each fit. Obtained R^2_{adj} values indicate that the tested objective metrics have a positive correlation with the listening test results ($p \leqslant 0.05$). Higher R^2_{adj} values are observed for different speakers across all tested objective metrics.

3 DISCUSSION

3.1 Authenticity vs. Transferring Task

The posterior contrast and the difference in the odds between the cases of using the same and different speakers indicate a significant perceptual difference caused by source type. While it was expected that detecting model-generated RIRs would be more challenging when source signals differ, the extent of this difficulty was previously unclear. Experimental results reveal that for some models, the increased difficulty with different speakers can alter conclusions about model fidelity.

For instance, given that the guessing rate in the 3-AFC task is $\frac{1}{3}$, `RT2RIR-LateOnly`, as shown in Fig. 6, would have been considered unsuitable for AR applications if only evaluated with the same speaker, where the MAP of the posterior distribution is 0.79. However, with different speakers, the MAP drops significantly to 0.48, suggesting that its perceived performance issues are less severe than initially thought. Similarly, the performance of `DarkVelvetNoise` and `FDN` approaches the guessing rate with different speakers, indicating that achieving an auditory illusion with these models is more feasible than it might have appeared based on evaluations conducted exclusively with the same speaker. These findings suggest that model fidelity may not need to be as high to meet the criteria for the transferring task as opposed to the authenticity task.

[2] https://github.com/TUIlmenauAMS/PsychoacousticLoss/tree/main.

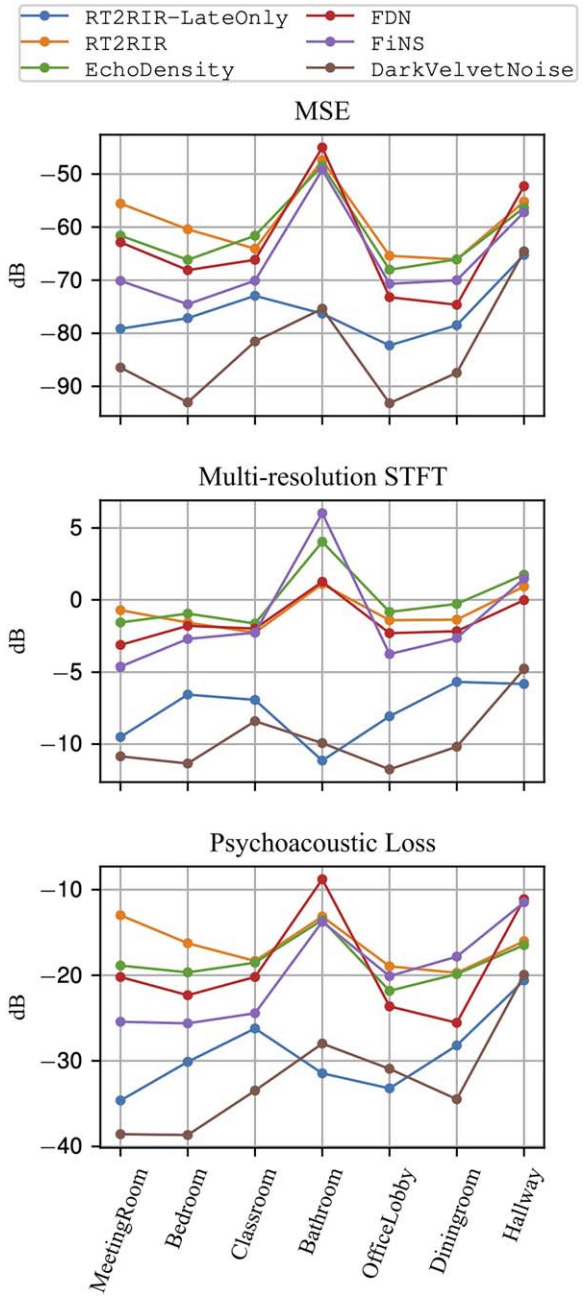


Fig. 7. Evaluation of each model per RIR using three objective metrics. Lower is better for all three metrics.

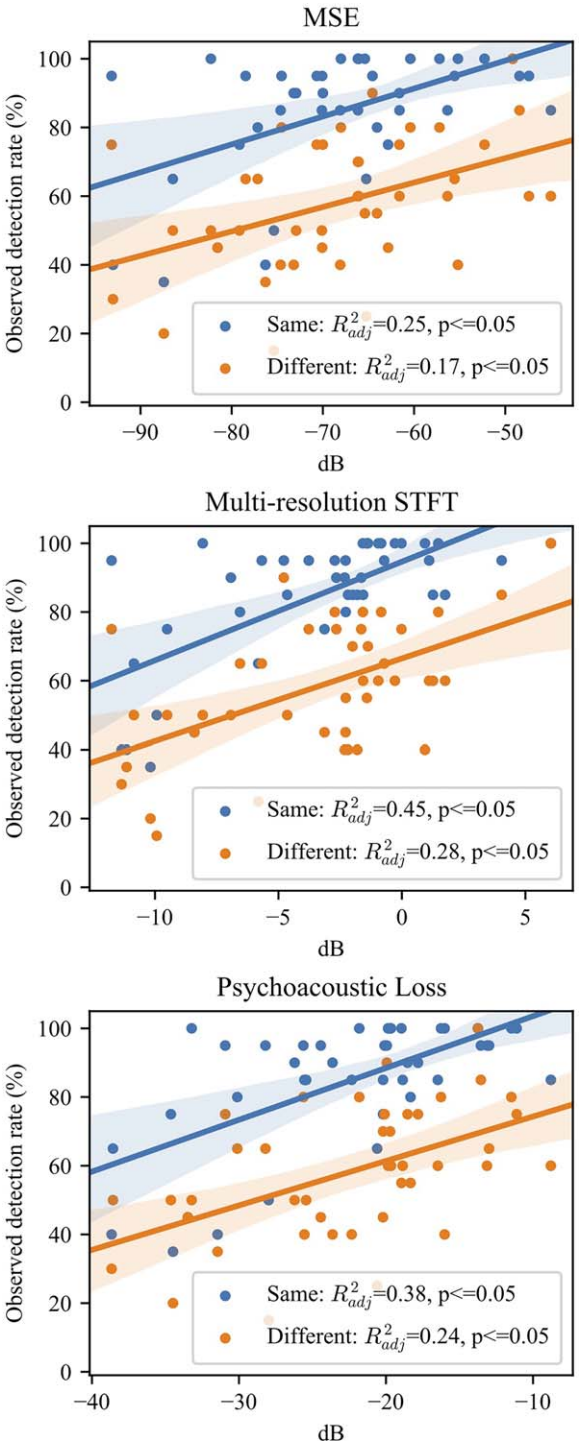


Fig. 8. Correlation between objective metrics and listening test results of using same and different speakers. R^2_{adj} values and p values of the linear least-squares regression results are shown in the legend. The objective metrics are on a decibel scale, and lower is better.

3.2 Listening Experience and Subject Acuity

There was no correlation between the average detection rate of each subject and their years of experience in acoustics research. Pearson correlation coefficients for the same and different speakers were $r = -0.08$ ($p \gg 0.05$) and $r = -0.20$ ($p \gg 0.05$), respectively. The variance in detection rates among subjects was small in trials with the same speaker but substantially higher in those with different speakers. This suggests that detecting with the same speaker was relatively easy for all subjects, though none achieved a perfect score, while different speakers posed greater difficulty with varying individual performance levels. Importantly, the results indicate that listening test performance was not influenced by expertise in acoustics, suggesting that the findings can be generalized to real-world users of AR applications.

Furthermore, the analysis of the posterior distribution of β parameters of the Subject predictor confirms that no subjects performed significantly better or worse than others.

Thus, it is concluded that subject acuity did not influence the overall results.

3.3 Accuracy of Early Reflection Modeling

Compared to late reverberation modeling, early reflections are more challenging to model due to their nonstatistical characteristics. Although it is unnecessary to model early reflections with sample-by-sample precision, capturing their salient aspects is still crucial. However, the precise level of accuracy required to achieve a plausible RIR remains uncertain, and this uncertainty is compounded when comparing RIRs with different speakers.

The primary difference between `RT2RIR-LateOnly` and `RT2RIR` lies in the accuracy of early reflections: `RT2RIR-LateOnly` directly copies the early reflections from the target RIR, whereas `RT2RIR` generates them from Gaussian noise. While this study employs a simplified approach to selecting the mixing time, as described in SEC. 1.3.1, a more sophisticated method could be explored in future assessments. Nevertheless, the current analysis is sufficient to demonstrate the impact of accurate versus inaccurate early reflection modeling.

By comparing the posterior mean of the model competency variable for `RT2RIR-LateOnly` and `RT2RIR` (Table 2), it is observed that accurate early reflection modeling has a significant effect. In the case of `RT2RIR`, the odds of correctly detecting the generated RIR are 1.79, meaning that a subject is 1.79 times more likely to detect the model-generated RIR correctly than not. Conversely, for `RT2RIR-LateOnly`, the odds drop to 0.62, indicating that a subject is less likely to detect the model-generated RIR correctly. This implies that inaccurate modeling of early reflections increases the odds of detection by a factor of 2.89. These findings highlight the important role of early reflection modeling in improving the detectability of model-generated RIRs, answering RQ3.

3.4 Listening Test vs. Objective Metrics

As shown in Fig. 8, positive correlations indicate that the metrics provide a first approximation of the perceptual quality; models that perform well on objective metrics are expected to perform well in the listening test. Among the tested objective metrics, the multiresolution STFT exhibits the highest R^2_{adj} values. Across all the tested metrics, the lower R^2_{adj} values are observed for different speakers, implying the limitation of these metrics in explaining the detection rate variability arising from source variations. Notably, none of the tested objective metrics account for comparisons between two reverberant signals with different speech content. This limitation is further reflected in the large spread observed in the mid-quality region, suggesting that the tested metrics are unable to fully capture the nuances required for comprehensive evaluation of reverberation models.

Furthermore, the absence of an equivalent to the "guessing rate" from the listening test complicates the interpretation of objective error values. Therefore, relying on these objective evaluations is insufficient to assess whether a model performs adequately for authenticity or transferring tasks. Developing objective metrics that can more effectively account for RIRs while incorporating source variability presents an interesting direction for improving the objective evaluation of reverberation models in AR applications.

4 CONCLUSION

This work compared the perceptual fidelity of reverberation models under the authenticity and transferring tasks with processed speech signals. The two tasks differed in source type. In the authenticity task, all three signals within each trial were spoken by the same speaker. In contrast, the transferring task featured three signals spoken by different speakers within each trial, ensuring greater variation in source characteristics.

The Bayesian statistical analysis revealed that the odds of detecting the model-generated RIR were 7.44 times higher with the same speaker than with different speakers. This suggests that designing tests with the same speaker is unsuitable for evaluating model performance, as it artificially increases detectability, especially when using different speakers better reflects real-world AR applications. Additionally, it was found that inaccurate modeling of early reflections increased the odds of detection by a factor of 2.88, underscoring the importance of early reflection fidelity in reverberation modeling. The impact of room acoustics on detection rate was generally small, except for one room, where detection was slightly easier. Similarly, subject acuity had no significant effect, as indicated by the 94% HPDI of all subjects' posterior distributions including zero, indicating similar performance levels across subjects.

Furthermore, a linear regression analysis of the listening test results and three commonly used objective metrics showed positive correlations. However, these metrics do not sufficiently distinguish between model-induced errors and variations due to source differences, limiting their reliability in assessing reverberation models.

Based on these findings, the authors recommend using the transferring task when evaluating static reverberation models, as it provides a more realistic and practical assessment. The insights gained from such evaluations can then contribute to the development of dynamic reverberation models for AR applications.

5 ACKNOWLEDGMENT

The authors thank Jon Fagerström and Gloria Dal Santo for providing their implementations of `DarkVelvetNoise` and `FDN` models.

6 REFERENCES

[1] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty Years of Artificial Reverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 5, pp. 1421–1448 (2012 Jul.). https://doi.org/10.1109/TASL.2012.2189567.

[2] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization—An Overview," *J. Audio Eng. Soc.*, vol. 41, no. 11, pp. 861–875 (1993 Nov.).

[3] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating Interactive Virtual Acoustic Environments," *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705 (1999 Sep.).

[4] C. Yeoward, R. Shukla, R. Stewart, M. Sandler, and J. D. Reiss, "Real-Time Binaural Room Modelling for Augmented Reality Applications," *J. Audio Eng. Soc.*, vol. 69, no. 11, pp. 818–833 (2021 Nov.). https://doi.org/10.17743/jaes.2021.0046.

[5] J. Yang, A. Barde, and M. Billinghurst, "Audio Augmented Reality: A Systematic Review of Technologies, Applications, and Future Research Directions," *J. Audio Eng. Soc.*, vol. 70, no. 10, pp. 788–809 (2022 Oct.). https://doi.org/10.17743/jaes.2022.0048.

[6] V. Välimäki, B. Holm-Rasmussen, B. Alary, and H.-M. Lehtonen, "Late Reverberation Synthesis Using Filtered Velvet Noise," *Appl. Sci.*, vol. 7, no. 5, paper 483 (2017 May). https://doi.org/10.3390/app7050483.

[7] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered Noise Shaping for Time Domain Room Impulse Response Estimation From Reverberant Speech," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 221–225 (New Paltz, NY) (2021 Oct.). https://doi.org/10.1109/WASPAA52581.2021.9632680.

[8] C. C. Darabundit, J. S. Abel, and W. Woszczyk, "Probabilistic Reverberation Model Based on Echo Density and Kurtosis," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 252–259 (Copenhagen, Denmark) (2023 Sep.).

[9] G. Dal Santo, K. Prawda, S. Schlecht, and V. Välimäki, "Differentiable Feedback Delay Network for Colorless Reverberation," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 244–251 (Copenhagen, Denmark) (2023 Sep.).

[10] ITU-B, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *ITU-R Recommendation BS.1534-3* (2015 Oct.).

[11] A. Lindau and S. Weinzierl, "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acust. united Acust.*, vol. 98, no. 5, pp. 804–810 (2012 Sep.), https://doi.org/10.3813/AAA.918562.

[12] S. A. Wirler, N. Meyer-Kahlen, and S. J. Schlecht, "Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes," in *Proceedings of the Audio Engineering Society International Conference on Audio for Virtual and Augmented Reality* (2020 Aug.), paper 3-4.

[13] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the Authenticity of Individual Dynamic Binaural Synthesis," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1784–1795 (2017 Oct.). https://doi.org/10.1121/1.5005606.

[14] N. Meyer-Kahlen, S. Schlecht, S. V. A. Garí, and T. Lokki, "Testing Auditory Illusions in Augmented Reality: Plausibility, Transfer-Plausibility, and Authenticity," *J. Audio Eng. Soc.*, vol. 72, no. 11, pp. 797–812 (2024 Nov.). https://doi.org/10.17743/jaes.2022.0178.

[15] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A Round Robin on Room Acoustical Simulation and Auralization," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2746–2760 (2019 Apr.). http://dx.doi.org/10.1121/1.5096178.

[16] T. McKenzie, N. Meyer-Kahlen, and S. J. Schlecht, "The Role of Source Signal Similarity in Distinguishing Between Different Positions in a Room," in *Proceedings of the AES International Conference on Spatial and Immersive Audio* (2023 Aug.), paper 25.

[17] N. Meyer-Kahlen, S. V. A. Garí, I. Ananthabhotla, and P. Calamia, "A Two-Dimensional Threshold Test for Reverberation Time and Direct-to-Reverberant Ratio," in *Proceedings of the Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, pp. 1–8 (Bologna, Italy) (2023 Sep.).

[18] A. Kuusinen and T. Lokki, "Recognizing Individual Concert Halls is Difficult When Listening to the Acoustics With Different Musical Passages," *J. Acoust. Soc. Am.*, vol. 148, no. 3, pp. 1380–1390 (2020 Sep.). https://doi.org/10.1121/10.0001915.

[19] D. Fantini, G. Presti, M. Geronazzo, et al., "Co-Immersion in Audio Augmented Virtuality: The Case Study of a Static and Approximated Late Reverberation Algorithm," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 11, pp. 4472–4482 (2023 Nov.). https://doi.org/10.1109/TVCG.2023.3320213.

[20] C. Schneiderwind, M. Richter, N. Merten, and A. Neidhardt, "Effects of Modified Late Reverberation on Audio-Visual Plausibility and Externalization in AR," in *Proceedings of the Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, pp. 1–9 (Bologna, Italy) (2023 Sep.).

[21] J. Traer and J. H. McDermott, "Statistics of Natural Reverberation Enable Perceptual Separation of Sound and Space," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 48, pp. E7856–E7865 (2016 Nov.). https://doi.org/10.1073/pnas.1612524113.

[22] A. Hoffmann, N. Meyer-Kahlen, and S. J. Schlecht, "Audiovisual Congruence and Localization Performance in Virtual Reality: 3D Loudspeaker Model vs. Human Avatar," *J. Audio Eng. Soc.*, vol. 72, no. 10, pp. 679–690 (2024 Oct.). https://doi.org/10.17743/jaes.2022.0162.

[23] G. Götz, R. Falcón Pérez, S. J. Schlecht, and V. Pulkki, "Neural Network for Multi-Exponential Sound Energy Decay Analysis," *J. Acoust. Soc. Am.*, vol. 152, no. 2, pp. 942–953 (2022 Aug.).

[24] G. Götz, S. J. Schlecht, and V. Pulkki, "Common-Slope Modeling of Late Reverberation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, no. 11, pp. 4472–4482 (2023 Nov.). https://doi.org/10.1109/TASLP.2023.3317572.

[25] J. S. Abel and P. Huang, "A Simple, Robust Measure of Reverberation Echo Density," presented at the *121st Convention of the Audio Engineering Society* (2006 Oct.), paper 6985.

[26] P. Huang and J. S. Abel, "Aspects of Reverberation Echo Density," presented at the *123rd Convention of the Audio Engineering Society* (2007 Oct.), paper 7163.

[27] K. Kawczinski and J. S. Abel, "Perceptually-Informed Features for Room Reverberation Modeling and Prediction," presented at the *149th Convention of the Audio Engineering Society* (2020 Oct.), paper 10446.

[28] M. A. Gerzon, "Synthetic Stereo Reverberation: Part One," *Studio Sound*, vol. 13, no. 12, pp. 632–635 (1971 Dec.).

[29] J.-M. Jot and A. Chaigne, "Digital Delay Networks for Designing Artificial Reverberators," presented at the *90th Convention of the Audio Engineering Society* (1991 Feb.), paper 3030.

[30] C. Kirsch, T. Wendt, S. van de Par, H. Hu, and S. D. Ewert, "Computationally-Efficient Simulation of Late Reverberation for Inhomogeneous Boundary Conditions and Coupled Rooms," *J. Audio Eng. Soc.*, vol. 71, no. 4, pp. 186–201 (2023 Apr.). https://doi.org/10.17743/jaes.2022.0053.

[31] J. Herre and S. Disch, "MPEG-I Immersive Audio—Reference Model for the Virtual/Augmented Reality Audio Standard," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 229–240 (2023 May). https://doi.org/10.17743/jaes.2022.0074.

[32] S. Schlecht, "FDNTB: The Feedback Delay Network Toolbox," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 211–218 (Vienna, Austria) (2020 Sep.).

[33] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks With Multi-Resolution Spectrogram," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203 (Barcelona, Spain) (2020 May). https://doi.org/10.1109/ICASSP40776.2020.9053795.

[34] S. V. Amengual Garí, J. M. Arend, P. T. Calamia, and P. W. Robinson, "Optimizations of the Spatial Decomposition Method for Binaural Reproduction," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 959–976 (2021 Dec.). https://doi.org/10.17743/jaes.2020.0063.

[35] K. Prawda, S. J. Schlecht, and V. Välimäki, "Dataset of Impulse Responses From Variable Acoustics Room Arni at Aalto Acoustic Lab," *Zenodo* (2022 Aug.). https://doi.org/10.5281/zenodo.6985104.

[36] M. Karjalainen and H. Järveläinen, "Reverberation Modeling Using Velvet Noise," in *Proceedings of the AES 30th International Conference: Intelligent Audio Environments* (2007 Mar.), paper 7.

[37] J. Fagerström, N. Meyer-Kahlen, S. J. Schlecht, and V. Välimäki, "Dark Velvet Noise," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 192–199 (Vienna, Austria) (2022 Sep.).

[38] J. Fagerström, S. J. Schlecht, and V. Välimäki, "Non-Exponential Reverberation Modeling Using Dark Velvet Noise," *J. Audio Eng. Soc.*, vol. 72, no. 6, pp. 370–382 (2024 Jun.). https://doi.org/10.17743/jaes.2022.0138.

[39] R. McElreath, *Statistical Rethinking: A Bayesian Course With Examples in R and Stan* (CRC Press, Boca Raton, FL, 2020).

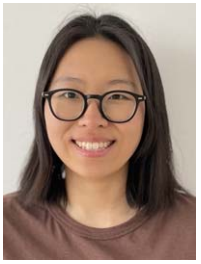
[40] R. Van de Schoot, S. Depaoli, R. King, et al., "Bayesian Statistics and Modelling," *Nat. Rev. Methods Primers*, vol. 1, no. 1, paper 1 (2021 Jan.).

[41] O. Abril-Pla, V. Andreani, C. Carroll, et al., "PyMC: A Modern, and Comprehensive Probabilistic Programming Framework in Python," *PeerJ Comput. Sci.*, vol. 9, paper e1516 (2023 Sep.).

[42] G. Schuller, *Filter Banks and Audio Coding: Compressing Audio Signals Using Python* (Springer, Cham, Switzerland, 2020), 1st ed.

[43] C. J. Steinmetz and J. D. Reiss, "Auraloss: Audio Focused Loss Functions in PyTorch," presented at the *Digital Music Research Network One-Day Workshop (DMRN+15)* (London, UK) (2020 Dec.).

THE AUTHORS


Kyung Yun Lee


Nils Meyer-Kahlen


Sebastian J. Schlecht


Vesa Välimäki

Kyung Yun Lee is a doctoral researcher at the Acoustics Lab in the Department of Information and Communications Engineering at Aalto University, Finland. She completed both her B.Sc. in Computer Science and M.Sc. in the Graduate School of Culture Technology at Korea Advanced Institute of Science and Technology (KAIST), South Korea. Her research interest is on developing perception models for audio for augmented reality using Bayesian statistics.

•

Nils Meyer-Kahlen is a postdoctoral researcher at the Acoustics Lab in the Department of Information and Communications Engineering at Aalto University, Finland, where he defended his D.Sc. in 2024. Before moving to Finland in 2019, he completed his B.Sc. and M.Sc. degrees in Electrical Engineering and Audio Engineering at the Technical University and the University of Music and Performing Arts in Graz, Austria. He interned as a research scientist at Reality Labs Research in 2022 and 2024. His research focuses on virtual acoustics for augmented reality, keeping technological development and human perception in sight.

•

Sebastian J. Schlecht is an Associate Professor of Signal Processing at the Chair of Multimedia Communications and Signal Processing at the University of Erlangen-Nuremberg, Germany. He earned a Diploma in Applied Mathematics from the University of Trier, Germany, in 2010, followed by an M.Sc. in Digital Music Processing at Queen Mary University of London, UK, in 2011. He completed his Doctorate at the International Audio Laboratories Erlangen, Germany, in 2017. From 2012 to 2019, he also served as an external research and development consultant at Fraunhofer IIS, Erlangen, Germany. Between 2019 and 2024, he held the position of Professor of Practice for Sound in Virtual Reality at Aalto University, Finland. He has received multiple best paper awards, including those from *JAES*, WASPAA, DAFX (2018, 2021, 2022, 2023), and AES AVAR.

•

Vesa Välimäki is a Full Professor of Audio Signal Processing and Vice Dean for Research at Aalto University, Espoo, Finland. He received his D.Sc. degree in electrical engineering from the Helsinki University of Technology in 1995. In 1996, he was a postdoctoral researcher with the University of Westminster, London, UK. From 2008 to 2009, he was a Visiting Scholar with Stanford University, Stanford, CA. He is a Fellow of the AES, the Institute of Electrical and Electronics Engineers (IEEE), and the Asia-Pacific Artificial Intelligence Association (AAIA). Prof. Välimäki was the Editor-in-Chief of the *Journal of the Audio Engineering Society* from September 2020 until September 2025.