Audio Engineering Society

Conference Paper

Presented at the Conference on
Immersive and Interactive Audio
2019 March 27 – 29, York, UK



# Evaluation of real-time sound propagation engines in a virtual reality framework

Sebastià V. Amengual Garí[1], Carl Schissler[1], Ravish Mehra[1], Shawn Featherly[1], and Philip W. Robinson[1]

[1] *Facebook Reality Labs*

Correspondence should be addressed to Sebastià V. Amengual Garí (sebastia.gari@oculus.com)

**ABSTRACT**

Sound propagation in an enclosed space is a combination of several wave phenomena, such as direct sound, specular reflections, scattering, diffraction, or air absorption, among others. Achieving realistic and immersive audio in games and virtual reality (VR) requires real-time modeling of these phenomena. Given that it is not clear which of the sound propagation aspects are perceptually more relevant in VR scenarios, an objective and perceptual comparison is conducted between two different approaches: one based on rendering only specular reflections of a geometrically simplified room (image source model - ISM), and another one based on ray-tracing using custom geometries. The objective comparison analyzes the simulation results of these engines and compare them with those of a room acoustic modeling commercial software package (Odeon), commonly employed in auralization of room acoustics. The perceptual evaluation is implemented in an immersive VR framework, where subjects are asked to compare the audio rendering approaches in an ecologically valid environment. In addition, this framework allows systematic perceptual experiments by rapidly modifying the test paradigm and the virtual scenes. The results suggest that the engine based on ISM is subjectively more preferred in small to medium rooms, while large reverberant spaces are more accurately rendered using a ray-tracing approach. Thus, a combination of both methods could represent a more appropriate approach to a larger variety of rooms.

## 1 Introduction

Compelling audio effects are a crucial element in ensuring an engaging experience in virtual reality (VR). High quality spatial audio capture and reproduction for linear non-interactive content, such as films or narrative experiences, can be achieved by using recorded audio signals. However, VR games are commonly highly interactive e.g. the user can navigate a space, engage in conversations with other users or AI characters, or generate sounds with their actions — and thus, spatial methods based on measurements are rarely flexible enough to provide a satisfactory experience. Sound synthesis techniques or libraries of pre-recorded sounds can be used to generate the sound emitted by the sources. However, since the sounds are radiated prior to arriving at the listener, they are heavily modified by the sound propagation properties of the scene, including specular and diffuse reflections, as well as diffraction, occlusion and/or transmission. Although in some cases, an audio engine can pre-compute ("bake") parts of the impulse response to reduce the computational cost of the simulation, a great portion of the sound propagation computation needs to be done in real-time. This leads

to physical simplifications and computational optimizations in order to comply with the time and compute budget constraints.

This paper presents an objective and perceptual evaluation of two sound propagation engines used for the real-time rendering of sound propagation effects in a virtual reality context. The objective evaluation is based on simulating room impulse responses with both engines for fixed source positions and listener orientations, and comparing them with the results of Odeon, a commercial offline room acoustic simulation engine [1]. For the perceptual simulation, a flexible interactive framework based on Unity has been implemented. Users are presented with virtual scenes and are asked to rate the audio renderings.

## 2 Sound propagation

The modeling of sound propagation can be separated in three main stages: source, medium, and receiver.

Modeling the source includes the implementation of frequency dependent directivity and level dependent power spectrum. The medium (or room) can be regarded as a linear filter applied on the sound generated by the source. However, this filter must model appropriately several physical phenomena such as specular reflections, scattering, diffraction, room modes, air absorption, sound transmission, and occlusion. Also, moving sources and listeners cause the presence of the Doppler effect, a change in the perceived frequency of the sound. While all these effects are governed by the wave equation and solving it in real time provides a physically accurate approach to the implementation of sound propagation effects, wave based simulations are usually computationally too expensive to be implemented in real-time [2]. Hence, given that above the Schroeder frequency the different wave fronts present in a room can be regarded as rays [3], alternative approaches based on geometrical acoustics (GA) are usually preferred for real-time applications [4]. However, GA simulations do not provide an accurate result at low frequencies, and thus hybrid models combining wave based approaches at low frequencies and GA at high frequencies represent a balanced approach between computational load and physical accuracy.

Finally, in AR/VR and game audio applications, the receiver is a human listener. Thus, the direct sound and propagation phenomena need to be properly spatialized and rendered in binaural audio format. To this end, the direct sound and propagation phenomena are appropriately filtered using a dataset of head-related transfer functions (HRTFs).

## 3 Audio engines

Game audio engines need to handle the spatialization and propagation effects of multiple sources in real-time, and the amount of compute available for this is usually fairly small. In addition, as opposed to architectural acoustic simulations, in game audio the scenes are typically dynamic, with moving sources, listener and geometry. Thus, to handle all these requirements in real time, computationally efficient approaches need to be used. For this reason, game audio engines have historically been based on the use of multiple reverberators controlled by low-level parameters which define directly the properties of the generated reverb e.g. decay time, low/mid/high frequency gain, or echo density, among others. In addition, other sound propagation effects such as transmission, occlusion, air absorption or diffraction are generally modeled using approximations based on low-pass filters [5, 6]. While these approaches can indeed produce perceptually satisfactory results, the final result depends on the technical skills and attention to detail of audio programmers and sound designers.

More recently, and with increasing available computing power, a number of game audio engines inspired by architectural acoustic simulation approaches have been developed. Examples of these are game audio engines that implement real-time geometrical acoustics (GA) [7] or partially pre-computed wave-based methods [8]. This approach imposes a change of paradigm in the audio design of game environments, given that a higher layer of abstraction is added between the design of the scene and the produced audio result. In this case, instead of directly designing multiple reverberators, the sound designer defines the geometry and material properties of the scene, and the audio engine then uses this information to render the resulting audio. The biggest advantage of these new approaches is the ability to handle large dynamic scenes with multiple environments and sources without the manual effort required by traditional game audio engines.

| Simulation aspect | Engine A | Engine B |
|---|---|---|
| HRTF dataset | CIPIC subject 48 | CIPIC subject 48 |
| HRTF equalization | Perception guided equalization. | Diffuse field equalization. |
| HRTF ITDs | ITDs extracted before SH expansion and reinserted into the binaural signal. | ITDs encoded into SH expansion. |
| Room Geometry | Shoebox | Fully customizable |
| Early Reflections | Shoebox model with listener fixed at the center | Raytracing with high diffusity. The early reflections are not prominent. |
| Late Reverb | Sampled RIRs. Static reverberation. | Raytracing, fully dynamic |
| Material absorption | Broadband, one material per wall (6 walls) | Customizable in 4 frequency bands |
| Material Scattering | No | Customizable in 4 frequency bands |
| Air absorption | No | Yes |
| Occlusion | No | Yes |
| Diffraction | No | No |

**Table 1:** Features of the evaluated sound propagation simulation engine.

### 3.1 Evaluated engines

The two real-time audio simulation engines evaluated in this paper implement approaches based on geometrical acoustics (GA), although simplifications are made due to computational limitations. Both engines are capable of generating spatialized sounds and propagation effects in real-time. They are compiled as Unity packages to allow a comparison in a flexible and interactive framework. A summary of the main features of the compared engines is included in Table 1.

*Engine A* is based on a simplified image source model (ISM) [9] of a shoebox room. The generated reflections correspond to a listener placed at the center of the room, and reflections are updated according to the head rotations of the listener. The absorption characteristics of each wall can be customized using a broadband value, and the direct-to-reverberant ratio can be arbitrarily modified. This engine can be regarded as an intermediate step between traditional and fully dynamic physics based game engines, given that the sound propagation effects are modeled using a higher level description of the scene e.g. room size and materials, instead of reverberation parameters - but multiple rooms or environments need to be explicitly defined and modeled as separated reverb zones. Effects such as occlusion, diffraction or air absorption are not modeled.

*Engine B* is based on a ray tracing approach implemented in four frequency bands and supporting the use of custom dynamic geometries [10, 11]. Material parameters are assigned to each triangle and can be specified in four frequency bands (See Table 2) with control of absorption and scattering coefficients. The fact that custom dynamic geometries can be used implies that the engine is fully dynamic, and thus it is not necessary to manually define different environments or reverb zones. In addition, effects such as occlusion or air absorption are implemented in the ray tracing simulation.

| Frequency band | Low frequency | High frequency |
|---|---|---|
| 1 | 0 | 176 |
| 2 | 176 | 775 |
| 3 | 775 | 3408 |
| 4 | 3408 | 20000 |

**Table 2:** Material and air absorption frequency bands in *Engine B*. Some overlap is present due to the 24 dB/octave crossover filters.

The HRTF dataset used by both engines is derived from a subject of the CIPIC database [12] and encoded in the spherical harmonics (SH) domain, although different post-processing has been applied. In *Engine A* the inter-aural time differences (ITD) of the head related impulse responses (HRIR) are extracted prior to the spherical harmonics expansion, and reconstructed at the end of the rendering pipeline. The final HRTF dataset is perceptually equalized using a sound design approach. For *Engine B*, the SH transform is performed on the complex HRTF, and diffuse equalization is applied.

## 4 Objective evaluation

The evaluated audio engines implement geometrical acoustics approaches, and thus aim at approximating the room impulse response (RIR) of a given scene. The objective evaluation consists of generating the binaural room impulse response (BRIR) of a set of shoebox

| L | H | W | $\alpha$ | Source position |
|---|---|---|---|---|
| 6 | 6 | 6 | 0.1 | (2.5, -2.5, 2.5) front, bottom, right |
| 6 | 6 | 6 | 0.2 | (2.5, -2.5, 2.5) front, bottom, right |
| 6 | 6 | 6 | 0.4 | (2.5, -2.5, 2.5) front, bottom, right |
| 6 | 6 | 6 | 0.6 | (2.5, -2.5, 2.5) front, bottom, right |
| 6 | 6 | 6 | 0.8 | (2.5, -2.5, 2.5) front, bottom, right |
| 6 | 6 | 10 | 0.1 | (2.5, -2.5, 4) front, bottom, right |
| 6 | 6 | 10 | 0.2 | (2.5, -2.5, 4) front, bottom, right |
| 6 | 6 | 10 | 0.4 | (2.5, -2.5, 4) front, bottom, right |
| 6 | 6 | 10 | 0.6 | (2.5, -2.5, 4) front, bottom, right |
| 6 | 6 | 10 | 0.8 | (2.5, -2.5, 4) front, bottom, right |
| 6 | 10 | 10 | 0.1 | (2.5, -4, 4) front, bottom, right |
| 6 | 10 | 10 | 0.2 | (2.5, -4, 4) front, bottom, right |
| 6 | 10 | 10 | 0.4 | (2.5, -4, 4) front, bottom, right |
| 6 | 10 | 10 | 0.6 | (2.5, -4, 4) front, bottom, right |
| 6 | 10 | 10 | 0.8 | (2.5, -4, 4) front, bottom, right |
| 10 | 10 | 10 | 0.1 | (4, -4, 4) front, bottom, right |
| 10 | 10 | 10 | 0.2 | (4, -4, 4) front, bottom, right |
| 10 | 10 | 10 | 0.4 | (4, -4, 4) front, bottom, right |
| 10 | 10 | 10 | 0.6 | (4, -4, 4) front, bottom, right |
| 10 | 10 | 10 | 0.8 | (4, -4, 4) front, bottom, right |

**Table 3:** Rooms modeled for the comparison of simulated RIRs.

rooms, and compare the results in terms of physiognomy of the pressure impulse responses and monaural room acoustic parameters. The same set of rooms is simulated as well using Odeon and its results are used as a reference for the comparisons. Odeon has been chosen as a reference because it is widely used for room auralization and it is based on hybrid geometrical acoustics (ISM and raytracing), thus it could be regarded as a quality target. The simulation parameters are the default parameters set by the 'Engineering' preset, with the number of late rays set to 50000 and a length of the RIR of 2 seconds. The details of the simulated rooms are summarized in Table 3.

### 4.1 Physiognomy of the RIR

An example of the left channel of a simulated BRIR is depicted in Fig. 1. It can be appreciated that the impulse response generated by *Engine A* contains only specular reflections, while *Engine B* presents a much more diffuse envelope with less prominent reflections. If the impulse responses are compared to Odeon, it appears that both engines render appropriately the temporal pattern of the RIR i.e. timing between strongest reflections. The direct-to-reverberant ratio of *Engine B* seems to be somewhat higher than the one of Odeon or *Engine A*, and a preliminary analysis comparing the energy decay curves (EDC) revealed that this trend is accentuated at high frequencies (4 kHz and above).

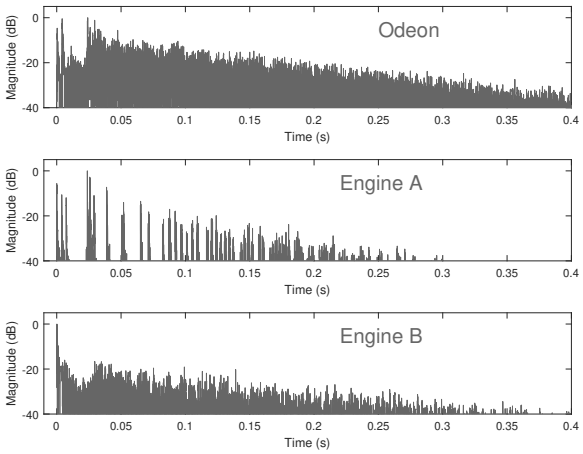


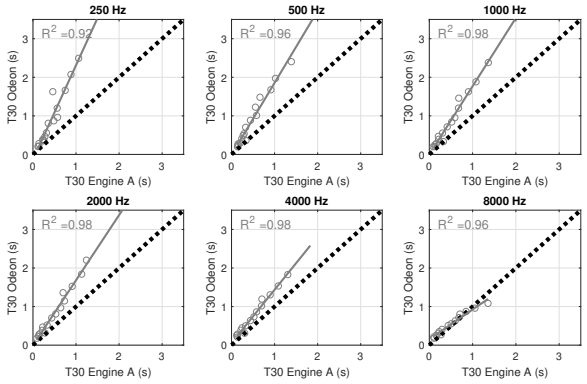

**Fig. 1:** Left channel of the simulated BRIR of a shoebox room with length, height and width of 10 m, and homogeneous absorption coefficient of $\alpha = 0.4$.

This could be attributed to the higher diffuseness of the responses generated by *Engine B*. Finally, the decay slope of *Engine A* is larger, resulting in a shorter and coarse reverberation tail.

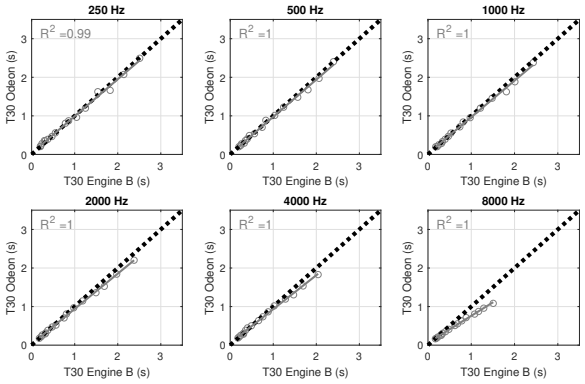
### 4.2 Room acoustic parameters

To further investigate the properties of the generated RIRs, standard monaural room acoustic parameters [13] i.e. Reverberation Time (T30), Early Decay Time (EDT) and Clarity (C50) are estimated using the left channel of the simulated BRIRs. Then, the results of *Engine A* and *Engine B* are mapped against those obtained from Odeon simulations. Figure 2 depicts the detailed results in octave bands. Results below 250 Hz are not included due to the known limitations of GA methods.

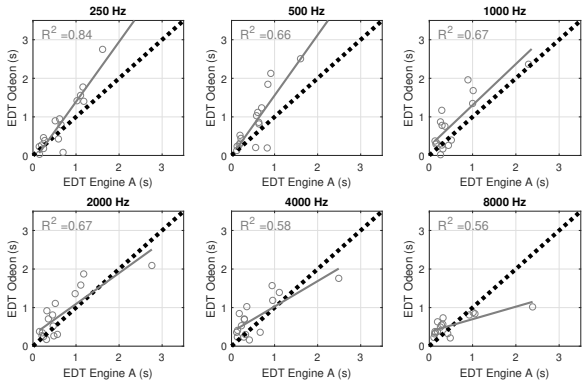
Both engines appear to render responses with a reverberation time (T30) that correlates strongly with the simulations from Odeon with values of adjusted $R^2$ higher than 0.9 for the frequency range from 250 Hz to 8 kHz. *Engine A* seems to usually underestimate the reverberation time, except at high frequencies - where the lack of modeling air absorption results in a better matched T30. For the case of *Engine B*, the reverberation time is closely matched and the differences are below the just noticeable difference (JND) (5% of the reference T30 [13]) in most of the cases. At high frequencies (8 kHz and above), the reverberation time is slightly overestimated. This is potentially due to



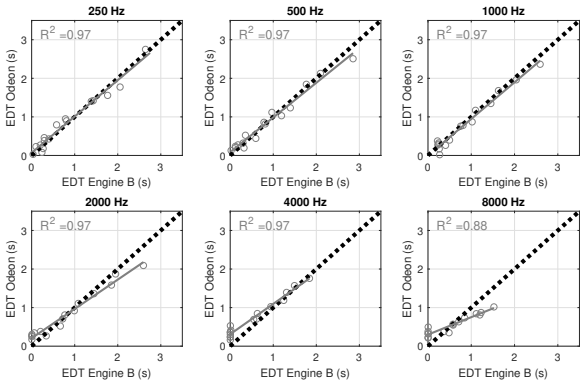

**(a)** T30 - Odeon vs Engine A



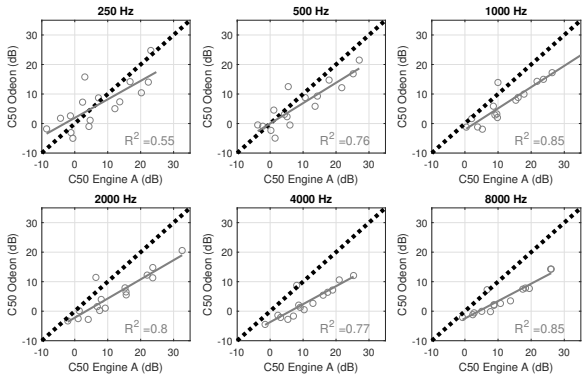

**(b)** T30 - Odeon vs Engine B



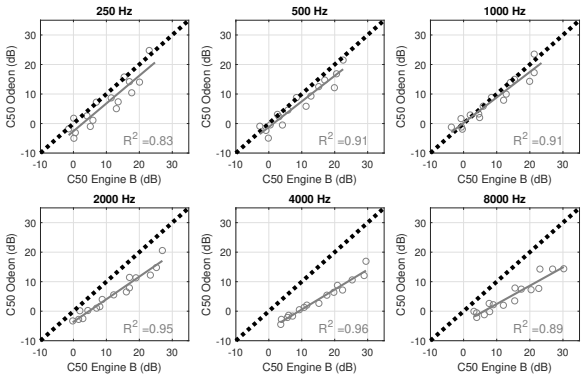

**(c)** EDT - Odeon vs Engine A




**(d)** EDT - Odeon vs Engine B




**(e)** C50 - Odeon vs Engine A




**(f)** C50 - Odeon vs Engine B

**Fig. 2:** Monaural parameters extracted from the BRIRs simulated with the evaluated game engines, linearly fitted with the results of Odeon 14.

having the same air absorption value for a fairly large frequency band (from 3.4 to 20 kHz - see Table 2).

With regards to the EDT, *Engine A* does not correlate as strongly as with T30. The EDT is usually underestimated at low frequencies and overestimated at high frequencies. In this case, given that the engine relies purely on an ISM, the initial decay of the RIR depends strongly on the timing and amplitude of the low order reflections. In the case of *Engine B*, the trend is much closer to the reference simulation, with slight deviations at high frequencies. Given that the early energy is much more diffuse, the early decay is not as abrupt as with *Engine A*, thus resulting in a more robust estimation.

Finally, the Clarity (C50) values of simulations corresponding to *Engine A* tend to correlate well for mid and high frequencies, although are generally overestimated. More variation is observed at lower frequencies. *Engine B* presents results of C50 that correlate well at all frequencies, and are fairly close to the reference in the range 250-1000 Hz. At higher frequencies the clarity is overestimated. For both engines, the deviations are typically considerably higher than the JND for C50 (1 dB [13]).

## 5 Perceptual evaluation

The perceptual evaluation consists of a listening test in VR. A variety of scenes have been selected from the Unity Asset Store, and their acoustics have been modeled using both engines. In order to have a flexible framework, both engines are compiled as Unity packages, and the listening test is implemented as a Unity 2017.3.1f1 project. This allows fast modifications of room geometry, sound sources, audio content and manipulation of material acoustic properties.

### 5.1 Modeled rooms

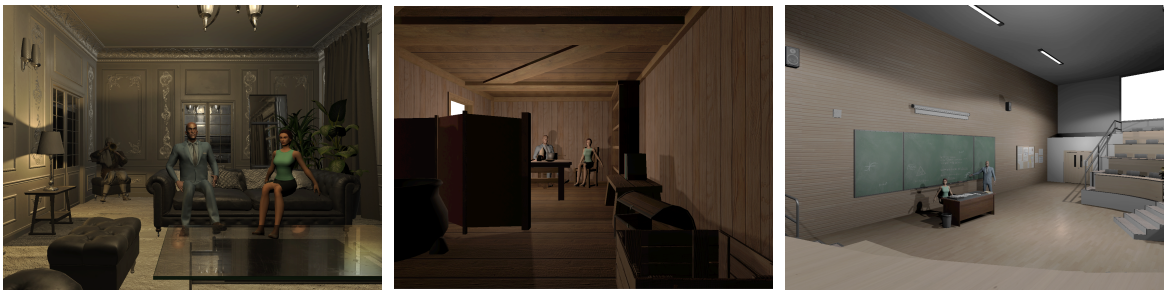
Six different rooms with varying room acoustic properties are included in the test. Images of the scenes are presented in Fig. 3. Each of the rooms has 3 sound sources placed at different positions: male and female speaker, and a trumpet player.

- *Living room* (T30 = 0.4 s): Shoebox scene with plaster walls and ceiling. Big carpet on the floor and absorptive furniture.
- *Cabin* (T30 = 0.7 s): Small wooden house with low absorption furniture and a highly absorptive room divider. The room has a second story and irregular ceiling height.
- *Lecture room* (T30 = 1.55 s): Large lecture room with wooden front wall and floor, plaster ceiling and curved back wall. Considerable amount of absorption at the audience area.
- *Warehouse* (T30 = 1.7 s): Large warehouse with concrete floor and filled with steel shelves which act as dividers between coupled corridors or subrooms providing many late reverberation paths. The effective volume of the room is considerably reduced by the presence of the shelves.
- *Space ship* (T30 = 2.5 s): Long room made out of steel with an opening at one end and inclined lateral walls.
- *Church* (T30 = 4.9 s): Gothic church with cross shaped floor plan and a rectangular main room. The only furniture present are wooden benches.

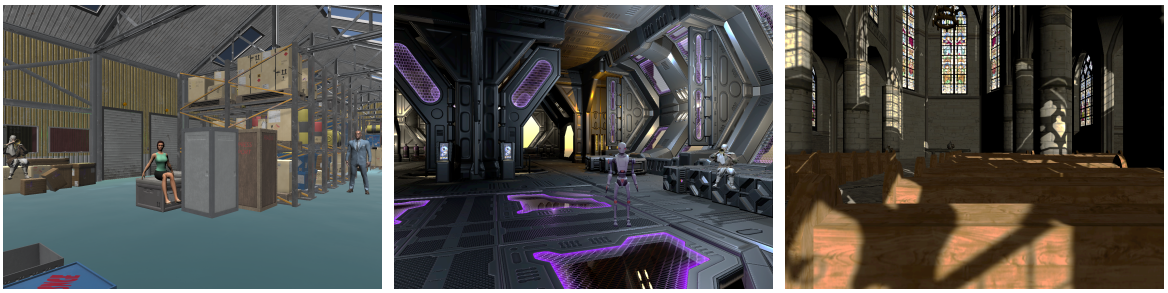
### 5.2 Matching room responses

As discussed in Section 4, the RIRs simulated using the evaluated engines differ considerably from each other, even when the room model is the same. For this reason, a procedure to match the responses of the rooms (to the best possible extent) has been applied. First, and provided that *Engine B* generates results that are closer to the reference (Odeon), the rooms are modeled using this engine. The absorption values used to model the materials are extracted from [14].

Having generated a scene with *Engine B*, the goal is now to generate comparable BRIRs using *Engine A*. The first step is to estimate the approximate total volume of the simulated room, and generate a shoebox model that gets as close as possible to the computed volume while maintaining the position of the main walls. This is done to obtain an early reflections pattern that matches with the overall geometry of the room. While this is quite straightforward for quasi-shoebox rooms — e.g. living room, it is more challenging for other scenes with curved walls or irregular geometries — e.g. lecture room (curved back wall), or wooden house (two story house with varying ceiling height). In these cases, the position of conflicting walls that cannot be easily modeled is chosen to result in the appropriate room volume.



**(a)** *Living Room* (T30 = 0.4 s) **(b)** *Cabin* (T30 = 0.7 s) **(c)** *Lecture Room* (T30 = 1.55 s)



**(d)** *Warehouse* (T30 = 1.7 s) **(e)** *Space Ship* (T30 = 2.5 s) **(f)** *Church* (T30 = 4.9 s)

**Fig. 3:** Room models used in the perceptual evaluation.

After modeling the geometry and total volume, appropriate material properties for *Engine A* must be selected. Given that the engine allows only broadband material absorption, it will be in practice not feasible to match the frequency-dependent properties of the reverberation. In addition, walls composed by several materials e.g. a wall with glass windows, cannot be modeled either. Thus, a single absorption value for each of the 6 walls in the shoebox model of *Engine A* is computed by averaging the materials of each wall.

After having modeled the geometry and absorption for *Engine B* and its simplified counterpart for *Engine A*, a validation of the RIRs is performed. This validation consists of measuring RIRs of the virtual scenes for positions of each of the three sources —male, female, and instrument. Then, the Energy Decay Curve (EDC), Reverberation Time (T20), Early Decay Time (EDT), and Clarity (C50) are estimated using the measured impulse responses. The target is to fit these parameters within approximately $\pm 2$ JND for the frequency bands between 250 Hz and 4000 Hz. Given that the responses generated by *Engine A* present reverberation times that are consistently lower than its theoretical equivalent, the material properties are modified until the target result is met. However, since the simulation approaches differ considerably, it is very likely that not all the parameters can be matched, specially EDT and C50, which rely on the very early part of the impulse response. In these cases the priority is given to match the slope and level of the EDC and the T20 values at the specified frequency bands. Matching the EDC often requires the modification of the level of the direct sound and early reflections with regards to the reverberation tail.

In order to render perceptually plausible scenes, the sound pressure level is calibrated. The sound pressure level of these sources is chosen to be approximately 60 dB SPL for male speech. Exceptions are a softer speaker used in the scene *Warehouse* (55 dB SPL) and a singing voice used in *Church* (66 dB SPL). Female speech is modeled to be 58 dB SPL and the trumpet is rendered at 80 dB SPL. To calibrate the presentation level, a head mounted display with headphones (Oculus Rift) is mounted on a calibrated measurement rig (MiniDSP E.A.R.S.) and the signals are reproduced in a free-field scene with a single source directly in front of the listener at 1 meter.

| Perceptual attribute | Description |
|---|---|
| Realism/Naturalness | The preferred renders resembles the better the expected sound and it sounds more natural. |
| Reverberance | The reverberation of one of the sounds is overall stronger (without specifying which one). |
| Reverberation quality | The reverberation of the selected sound has overall a better perceived quality. |
| Spatial impression | The spatial properties of the selected sound fit better to the spatial visual properties of the room (presence of reflections, envelopment of the sound, direction of echoes...). |
| Sound color | There is a difference in tonal balance between the presented sounds e.g. one of the sounds is darker. |
| Presence of artifacts/defects | The selected sound presents less artifacts or defects than the other one (or is free of artifacts). Artifacts are regarded as parasitic sound properties that shall not be present e.g. clicks, hissing noise, comb filter effect... |
| Source localization | The selected sound is localized closer to the ground truth source position e.g. mouth, trumpet. |
| Source distance | The distance of the presented sounds is different (without specifying which one is closer). |
| Source width | One of the sounds is bigger in size than the other one (without specifying which one is bigger). |
| Sound clarity | One of the sounds can be heard more clearly or is more intelligible (without specifying which one). |
| Loudness | One of the sounds is louder than the other one (without specifying which one is louder). |
| Sound externalization | The preferred sound is better externalized - perceived to be out of the listener's head. |

**Table 4:** Perceptual attributes used in the listening test.

### 5.3 Test procedure

The perceptual test consists of comparing the sound rendering produced by the two compared engines in multiple scenes with different visual and room acoustic properties. The virtual scenes are presented using a head-mounted display (Oculus Rift) and the test interaction is completed using VR controllers (Oculus Touch). To approximate the experimental conditions to real world conditions, the headset's built-in headphones are used, and no equalization is performed. Listeners are allowed to move within the available physical space (approximately 1.5 $\times$ 2 meters), and head rotations are allowed and encouraged. During each trial the listener can freely switch between the two renderings in real-time and listen to the scenes for as much time as desired.

The test is composed of 36 trials (6 rooms $\times$ 3 sources $\times$ 2 repetitions), and the task for each trial is divided in three parts:

1. Subjective preference: In a two-alternative forced choice (2AFC) task, the listener has to choose which audio rendering of the scene (*Engine A* or *Engine B*) fits better with their subjective expectations of the room, according to how they think the rooms should sound. In each of the trials the order of presentation is randomized, so listeners do not have explicit information about which of the renderings they are listening to.
2. Confidence: The subject is asked to rate in a 5-point Likert scale how confident they are about the answer to question 1. It is expected that some users will not have a clear internal reference, or that there are conflicting qualities in the sound renderings, and a compromise is made in order to answer question 1.
3. Perceptual attributes: The user is presented with a list of perceptual sound attributes and asked to select 3 attributes that they consider are most different between engines. A list of attributes and their definitions is summarized in Table 4. The meaning of these is discussed with the subjects before the test, during a short training period, ensuring a homogeneous understanding.

After the test is finished, subjects are asked to complete a short survey to generate population data statistics and provide free form feedback about the evaluated engines and the characteristics of the listening experiment.

Thirteen participants (12 male, 1 female) with normal hearing conditions (self-reported) completed the listening test. Among those subjects, seven are considered to be expert listeners (reported themselves as having 'a great deal' of familiarity with critical listening) and six of them reported a 'great deal' of familiarity with room acoustics terminology.

## 6 Perceptual experiment results

A graph summarizing the perceptual results is presented in Fig. 4. When grouping the results of all subjects and rooms, *Engine A* is preferred approximately 65% of the trials, and *Engine B* is preferred during the remaining

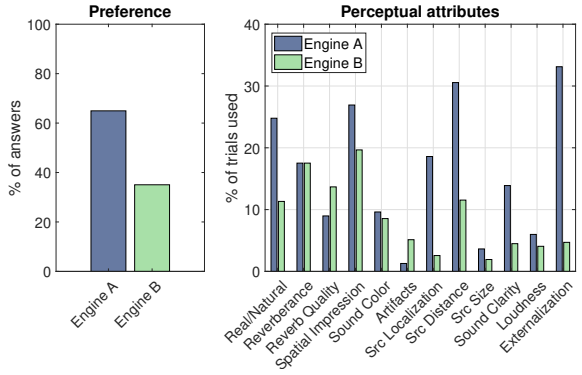



**Fig. 4:** Results of the perceptual experiment.

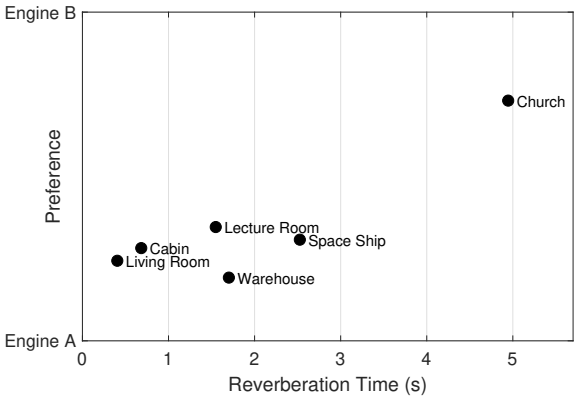



**Fig. 5:** Subjective preference mapped against reverberation time of the scenes.

35%. From the perceptual attribute selection, it is observed that *Engine A* tends to provide better perceptual results for *Realism/Naturalness*, *Spatial Impression*, *Source Distance* and *Externalization*. On the other hand, *Engine B* generally presents better *Reverberation Quality*. *Reverberance* presents the same number of appearances for both engines. A collection of free form comments made by the subjects is presented in Table 5. These reinforce the results from the formal test and reveal interesting information. For instance, when referring to *Engine B*, some subjects mentioned high frequency reverberation artifacts detrimental to their experience. However, these artifacts are positively appreciated in *Space Ship*, where they are regarded as a slightly metallic effect.

Given that the set of rooms used in the perceptual experiment present vastly different acoustic conditions, the same analysis is repeated for each of the rooms. Mapping the subjective preference against the reverberation time (T20) of the room (see Fig. 5) suggests that the subjective preference potentially depends on the acoustic conditions of the room. For small and mid sized rooms, *Engine A* (Image Source Model) seems to render better externalization and source distance by providing strong early reflections. However, in larger reverberant environments, *Engine B* (frequency dependent ray tracing) renders a smoother and perceptually more preferred reverberation tail. Thus, a combination of both engines could improve the perceptual impression in a wider set of rooms.

The tracking data reveals that although users are allowed and encouraged to move within the designated space, they tend to mainly perform rotational movements on the azimuth plane and stay at the same listening position.

## 7 Conclusion

This paper presents the objective and perceptual comparison of two real-time sound propagation engines. *Engine A* is based on an ISM, and *Engine B* is based on ray-tracing. The objective comparison reveals that *Engine B* produces results that are comparable to Odeon in terms of monaural room acoustic parameters at mid frequencies. *Engine A* is usually perceptually preferred for small to medium rooms, due to a better rendering of *Externalization, Source Localization*, and *Source Distance*. *Engine B* is generally preferred in bigger, more reverberant environments, due to its higher *Reverberation Quality*. A combination of both approaches could produce perceptually satisfactory results in a wider variety of rooms.

A common comment given by listeners in the perceptual test refers to the compromise that a 2AFC paradigm imposes on their decision when an engine performs better for some but not all of the perceptual attributes. Analyzing the subject consistency in their AFC responses show that subjects are consistent in average during 73% of the times, meaning that their response is not repeatable for a significant portion of the trials. Thus, in future tests, an alternative paradigm could be used to address this concern. In addition, quantitative scales e.g. sliders, could be added to rate the quality of the sound attributes.

The framework and test procedure presented in this paper allows for systematic perceptual evaluation of game

| | *Engine A* | *Engine B* |
|---|---|---|
| Positive | Externalization and localization are better.<br>Fast head tracking.<br>In *Living Room* the distance is rendered better.<br>Sharper high frequency response added realism for close sources.<br>In *Space Ship*, the slightly metallic speech fits better. | Deeper with slightly nicer coloration.<br>Artifact free reverberance in large rooms.<br>Less discrete reflections in small rooms.<br>It seemed to work better in small to medium spaces, but sometimes in big ones too.<br>Better spatial impression.<br>Smooth reverberation. |
| Negative | 'Slappy' sort of reverb.<br>Sharper high frequency response detracted from realism of far sources.<br>More artifacts.<br>Hissing noise in the church (*'sss'*). | Collapse/decrease in distance.<br>Low-passed direct sound detracted from realism in trumpet and female voice cases.<br>Poor externalization/Internalized sources.<br>Lag in spatial update led to localization instability.<br>More 'diffuse' localization in *Lecture Room*. Speech is perceived as coming from loudspeakers. |

**Table 5:** Verbal feedback from participants of the listening test.

audio engines. Furthermore, the scenes can be easily modified to result in more ecologically valid environments by adding ambient sound sources, generating more complex sonic scenes and giving the user the ability to teleport and explore a larger portion of the virtual scene.

## References


[1] Naylor, G. M., "ODEON - Another hybrid room acoustical model," *Applied Acoustics*, 38(2-4), pp. 131–143, 1993.

[2] Vorländer, M., "Computer simulations in room acoustics: Concepts and uncertainties," *The Journal of the Acoustical Society of America*, 133(3), pp. 1203–1213, 2013.

[3] Schroeder, M. R., "The "Schroeder frequency" revisited," *The Journal of the Acoustical Society of America*, 99(5), pp. 3240–3241, 1996, doi:10.1121/1.414868.

[4] 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.

[5] 3D Working Group of the Interactive Audio Special Interest Group, "Interactive 3D Audio Rendering Guidelines Level 2.0," Technical report, 1999.

[6] Hiebert, G., "OpenAL 1.1 Specification and Reference," Technical report, Creative Labs, Inc., 2005.

[7] Schissler, C. and Manocha, D., "GSound: Interactive Sound Propagation for Games," in *Audio Engineering Society Conference: 41st International Conference: Audio for Games*, 2011.

[8] Raghuvanshi, N., Snyder, J., Mehra, R., Lin, M., and Govindaraju, N., "Precomputed Wave Simulation for Real-time Sound Propagation of Dynamic Sources in Complex Scenes," *ACM Trans. Graph.*, 29(4), pp. 68:1–68:11, 2010, ISSN 0730-0301, doi:10.1145/1778765.1778805.

[9] 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.

[10] Schissler, C. and Manocha, D., "Adaptive impulse response modeling for interactive sound propagation," in *Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pp. 71–78, ACM, 2016.

[11] Schissler, C., Stirling, P., and Mehra, R., "Efficient construction of the spatial room impulse response," in *2017 IEEE Virtual Reality (VR)*, pp. 122–130, 2017, ISSN 2375-5334, doi:10.1109/VR.2017.7892239.

[12] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C., "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pp. 99–102, 2001, doi:10.1109/ASPAA.2001.969552.

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

[14] Egan, M., *Architectural Acoustics*, Classics (J. Ross), J. Ross Pub., 2007, ISBN 9781932159783.