# Fast Synthesis of Perceptually Adequate Room Impulse Responses from Ultrasonic Measurements

Jing Yang[1], Felix Pfreundtner[1], Amit Barde[2], Kurt Heutschi[3], Gábor Sörös[1]
[1]Department of Computer Science, ETH Zurich, Zurich, Switzerland
[2]Auckland Bioengineering Institute, The University of Auckland, Auckland, New Zealand
[3]Empa, Swiss Federal Laboratories for Materials Science and Technology, Dübendorf, Switzerland

## ABSTRACT

Audio augmented reality (AAR) applications need to render virtual sounds with acoustic effects that match the real environment of the user to create an experience with strong sense of presence. This audio rendering process can be formulated as the convolution between the dry sound signal and the room impulse response (IR) that covers the audible frequency spectrum (20Hz - 20kHz). While the IR can be pre-calculated in virtual reality (VR) scenes, AR applications need to continuously estimate it. We propose a method to synthesize room IRs based on the corresponding IR in the ultrasound frequency band (20kHz - 22kHz) and two parameters we propose in this paper: slope factor and RT60 ratio. We assess the synthesized IRs using common acoustic metrics and we conducted a user study to evaluate participants' perceptual similarity between the sounds rendered with the synthesized IR and with the recorded IR in different rooms. The method requires only a small number of pre-measurements in the environment to determine the synthesis parameters and it uses only inaudible signals at runtime for fast IR synthesis, making it well suited for interactive AAR applications.

## CCS CONCEPTS

• **Human-centered computing** → *Mixed / augmented reality*; *Auditory feedback*; • **Applied Computing** → *Sound and music computing*.

## KEYWORDS

Room acoustic effects, Room impulse response, Ultrasound, Augmented reality, Auditory perception

**ACM Reference Format:**
Jing Yang, Felix Pfreundtner, Amit Barde, Kurt Heutschi, Gábor Sörös. 2020. Fast Synthesis of Perceptually Adequate Room Impulse Responses from Ultrasonic Measurements. In *Proceedings of the 15th International Audio Mostly Conference (AM'20), September 15–17, 2020, Graz, Austria.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3411109.3412300

## 1 INTRODUCTION

When we listen to the same sound in different environments, our perception of the sound varies significantly due to the differing acoustic properties of these environments. For example, a voice in a large space with reflective surfaces such as a church will sound completely different in comparison to the same voice heard in a small carpeted room with furniture in it. In augmented reality (AR) and virtual reality (VR) applications, it is important that the rendered audio signals sound like they "belong" in the environment. This has been shown to improve a user's sense of presence and immersive experience [14].

A typical approach to capturing the "acoustic fingerprint" of a space is to measure the impulse response (IR) and then compute the convolution between the IR and the dry sound. To exactly replicate how a space sounds, IRs from every point in the space are supposed to be captured, because an IR is a function of room geometry, materials, and the positions of the source and the listener. However, carrying out IR measurements for every source-receiver location is laborious and time-consuming. To address this issue, some researchers simulate IRs by modeling the sound propagation process based on the 3D environment model and the acoustic properties of the indoor materials [8]. In recent years, thanks to the development in the field of computer vision, some researchers reconstruct the environment geometry and classify the surface materials using visual inputs from cameras [20]. Some researchers also optimize simulated IRs based on target acoustic parameters (e.g. reverberation time) that are estimated using deep learning models [23].

We present a method to synthesize IRs for arbitrary listener positions. Unlike existing methods that model a sound propagation process with detailed 3D geometry and material properties, our method is based on a set of IR-related parameters proposed in this work. Our method uses measurements in the ultrasound frequency band (20kHz - 22kHz, referred to as ultra-IR in the following) as input for "expansion" to the full IR that covers the audible frequency spectrum. Given an ultra-IR, we detect the arrival time of the direct sound, and use it to initialize IRs in each octave band (center frequency from 63Hz to 16kHz, referred to as octave-IR in the following) with an exponentially decaying Gaussian white noise. The early reflections and late reverberation parts of each initialized signal are then corrected to finalize the octave-IR. Finally, octave-IRs are added up and normalized to generate the room IR in the full frequency range. Rather than reconstructing a perfect room IR from ultra-IR, our goal is to create convincing acoustic effects with the well-enough approximated IR for AR applications.

The proposed method appears promising in the following aspects. First, this method does not require the simulation of the entire sound propagation process that relies on the environment geometry and the material information. Secondly, machine learning-based approaches usually require a large number of training samples, while we only need a small number of pre-recorded IRs (16 in


*AM'20, September 15–17, 2020, Graz, Austria*

ACM ISBN 978-1-4503-7563-4/20/09…$15.00
https://doi.org/10.1145/3411109.3412300

our examples) in the environment of interest to determine the IR synthesis parameters: slope factor and RT60 ratios. Furthermore, since the method requires to play and measure ultrasound signals that are inaudible to most people, it will not disturb the user at runtime of an application.

We evaluated the proposed method using both objective comparison and subjective user study. Objectively, we compared the synthesized IR with the ground truth IR in terms of common acoustic metrics: reverberation time, early decay time, clarity 80, and center time. Subjectively, we conducted a user study and found that participants generally leaned towards similar auditory perception between the sounds rendered with the real IR and with the synthesized IR from our method.

The main contributions of this work are: (1) a novel ultra-IR to room IR synthesis method that creates perceptually adequate acoustic effects for immersive auditory experience; (2) objective and subjective evaluations of this IR synthesis method.

The proposed method has high potential to be applied in interactive AR applications. The pre-computed room-dependent parameters can be stored in the cloud and retrieved for IR synthesis when the user enters the space. Moreover, this parametric synthesis approach runs significantly faster than modeling a complete sound propagation process, making it well suited for real-time implementation as the user moves around in the space.

## 2 RELATED WORK

Rendering a realistic acoustic environment in AR and VR applications is important to enhance a user's sense of presence [27] and has been a subject of research for several years [7, 21]. To this end, a typical approach is to dynamically compute the IR of an environment based on the sound propagation path in the space given a specific source-receiver position. This computation process usually requires the 3D environment geometry and the acoustic properties of the materials in the environment.

Hulusic et al. [8] provide a comprehensive overview of the major techniques for modeling the sound propagation process given the environment model and the material properties. Wave-based techniques like the boundary element method (BEM) [11] aim to solve the wave equations that describe the sound propagation process very accurately, but require long processing time. Techniques like the image source method (ISM) [2], volumetric methods [8], and particle-based methods [8] simplify the modeling process by excluding some properties of sound waves such as scattering and diffraction. The modeling results are not as accurate as the wave-based techniques, especially for some specific frequency bands, but the computational requirement is lower. There also exist GPU accelerated approaches [19, 24] and ray-based approaches [8] that significantly save the simulation time. Although results may demonstrate lower accuracy than the above approaches due to simplified acoustic approximation, these methods still generate perceptually satisfactory acoustic effects for AR and VR applications [10, 20].

In recent years, the development in the field of computer vision has motivated researchers to use visual inputs from cameras for 3D environment reconstruction [10, 15, 20] and material classification [9, 20]. Based on the modeled geometry and the classification of materials, the above sound propagation modeling approaches can be applied to generate the desired IRs. Alternatively, some researchers use the 3D model and the material classification as inputs for plugins that directly render the virtual sound with appropriate acoustic effects for VR users [9]. Despite these advances, the potential errors in the geometry modeling and material classification may lead to unsatisfactory rendering of the acoustic environment. To address this shortcoming, some researchers have attempted to integrate an IR optimization step. This optimization process might implement a solver system based on the ground truth IRs that have been pre-recorded in the environment [20], or train a deep learning network to estimate target acoustic parameters (e.g. reverberation time) in the environment and use these target parameters to optimize the sound rendering [23]. Deep learning-based approaches usually require a large number of training samples. Due to the difficulty of a large-scale in-situ IR measurement, researchers augment existing IR datasets before training a deep network [23].

In addition to the geometry and material-based sound propagation modeling methods, some researchers propose to statistically code a parametric sound wave field to simulate desired acoustic environments [17, 18]. According to Raghuvanshi and Snyder [17], much of the perceptual quality of a rendered sound can be captured by a few acoustic parameters of an IR (e.g. decay time of the late reverberation). Therefore, they propose a method that codes the field of time-varying IRs in terms of a few pre-computed parameters. This parametric method runs significantly faster than the sound propagation modeling approaches and it generates satisfactory acoustic effects for VR applications where a user's surroundings dynamically change with the user's movement.

Similar to Raghuvanshi and Snyder's concept [17], we intend to synthesize IRs for real-world environments based on several target parameters instead of modeling a complete sound propagation process. To determine the target parameters, our method requires a small number of pre-recorded IRs (around 20) in the real environment. Then, during the synthesis process, our method creates perceptually adequate acoustic effects by only using ultrasonic measurements as input to approximate a complete room IR for arbitrary listener positions.

## 3 FROM ULTRA-IR TO ROOM IR

In this section, we elaborate on the method of synthesizing a room IR from its corresponding ultra-IR. Our goal is to approximate a room IR that produces perceptually adequate acoustic effects rather than reconstructing a perfect, original IR. The method first synthesizes IRs in each octave band with center frequency $f_c \in$ [63Hz, 125Hz, 250Hz, 500Hz, 1kHz, 2kHz, 4kHz, 8kHz, 16kHz] based on the ultra-IR, and then add up the octave-IRs to generate the room IR. According to [12], it is common to segment an IR into three parts in acoustic analysis: direct sound (DS), early reflections (ER), and late reverberation (LR). As shown in Figure 1, the method of constructing each octave-IR consists of four steps considering these three parts: determining the arrival time of the DS, initializing each octave-IR with exponentially decaying Gaussian white noise (GWN), correcting the ER, and adjusting the LR. In this paper, the sampling rate of IRs is 44.1kHz unless noted otherwise.

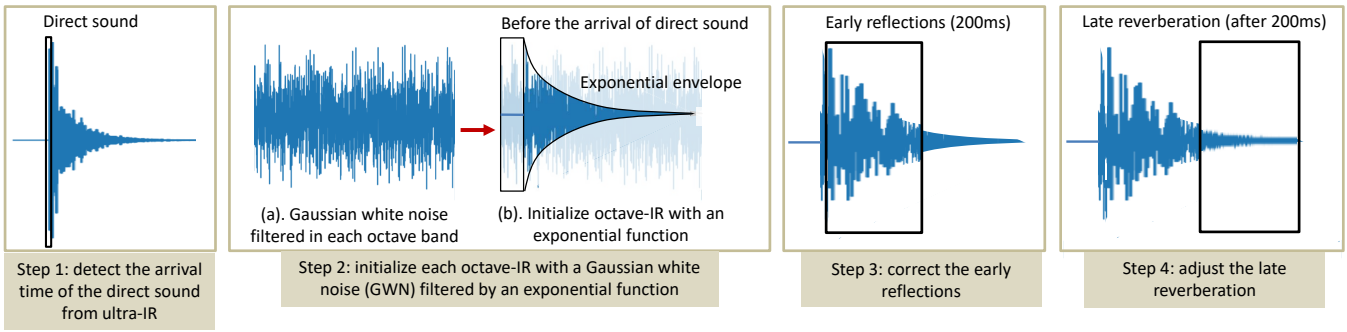


**Figure 1: Four steps to generate an octave-IR based on an ultra-IR: the arrival time of the direct sound is detected from ultra-IR and is used as the direct arrival time for each octave band (Step1). Given the direct arrival time, each octave-IR is initialized with a Gaussian white noise filtered by an exponential function (Step2). After initialization, the early reflections consisting of specular and diffuse components are corrected based on a target energy value (Step3). Finally, the late reverberation part is adjusted based on a target energy value (Step4). Normalized octave-IRs are added up and normalized to generate the room IR.**

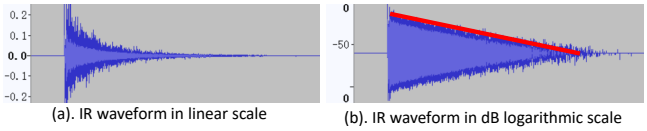


**Figure 2: The waveform of an IR shown in linear scale and in dB logarithmic scale. The envelope of the waveform can be approximated by a straight line in logarithmic scale.**

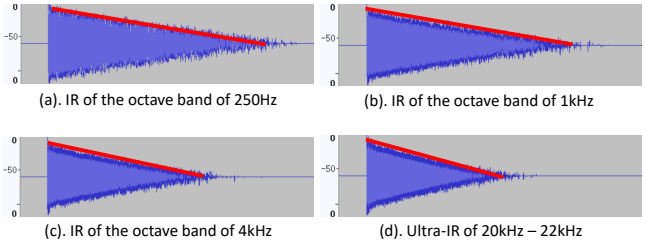


**Figure 3: When we split an IR into frequency bands, the slope of the envelope may be different in each frequency band.**

## 3.1 Arrival time of the Direct Sound

The arrival time of the direct sound indicates the onset of an IR. We detect the arrival time in the ultra-IR (20kHz - 22kHz) and use it as the onset time for each octave-IR.

The sample amplitudes $a_s$ in an ultra-IR signal are first normalized to the range $[-1, 1]$. To detect the arrival time of the direct sound, as commonly applied in audio signal processing [4, 13], we shift a window across the ultra-IR and calculate the energy of each segment $E(w) = \sum a_s^2$ as well as the derivative of the energy $\frac{\mathrm{d}E(w)}{\mathrm{d}w}$. We use a window size of 128 samples and a step size of 32 samples. The first window where the derivative peaks, and the energy of the window is above the set threshold, is selected as the window of the direct sound. An energy threshold of 0.6 works robustly in all our examples. We use the time-stamp of the first sample in the selected window as the arrival time of the direct sound.

## 3.2 Initialization of Octave-IR

After the arrival time of the direct sound is detected, our next step is to initialize an IR in each octave band. We will correct the ER and LR parts in each octave-IR after the initialization.

Inspired by [17], we propose to initialize each octave-IR with a Gaussian white noise (GWN) through the following processing steps. As shown in Step 2 in Figure 1, a GWN of the same length as the ultra-IR is first filtered into each octave band. The samples before the arrival of the direct sound are set to 0 and the samples in the direct arrival window remain unchanged. Then, an exponential function is applied on the GWN after the window of the direct sound, initializing the octave-IR with an exponential envelope.

The exponential function can be formulated according to the envelope of the target octave-IR in the dB logarithmic scale. As shown in Figure 2, the envelope of an IR in the dB logarithmic scale can be approximated by a straight line. Therefore, we need to determine the slope $S$ of this line, and then the IR envelope in the linear scale, i.e. the exponential function to filter the GWN, can be formulated as $E_{linear} = 10^{E_{dB}/20}$, where $E_{dB} = S * Index_{sample}$.

To determine the slope $S$ for the exponential function in each octave band, we assume that there exists a slope factor $f_{slope}$ that fulfills

$$f_{slope} = \frac{S_{ultra}}{S_{octave}} * \frac{RT60_{ultra}}{RT60_{octave}} \quad (1)$$

where $S_{ultra}$ and $S_{octave}$ represent the slope of the envelopes of the ultra-IR and each octave-IR in the dB logarithmic scale, and $RT60_{ultra}$ and $RT60_{octave}$ represent the reverberation time (RT60) of the ultra-IR and each octave-IR. We propose this assumption based on the fact that the decay of an IR varies across frequency bands (as shown in Figure 3), which is related to the frequency-dependent air attenuation and surface absorption in the environment, which leads to varying RT60 values in different frequency bands.

We suppose that $f_{slope}$ can be taken as a *room constant* and can be determined by applying Eq(1) with $S_{ultra}$, $S_{octave}$, $RT60_{ultra}$,

and $RT60_{octave}$ obtained from IRs that are pre-measured in the real environment. We also assume that the ratio $RT60_{ultra}/RT60_{octave}$ can be determined from the pre-measured IRs and used as an *octave constant* for each octave band. After the parameters $f_{slope}$ and $RT60_{ultra}/RT60_{octave}$ are determined, given a new input ultra-IR (i.e. a new $S_{ultra}$) that can be live captured at runtime, $S_{octave}$ can be calculated using Eq(1) and then used to determine the exponential function by $E_{linear} = 10^{S_{octave} * Index_{sample}/20}$ for initializing the octave-IR.

*3.2.1 Determining the Slope Factor and the RT60 Ratios in Simulated Rooms.* The above discussion brought us to exploring the slope factor and the RT60 ratios, and investigating whether these parameters generally exist in different rooms.

We first explored the parameters using IRs simulated by the room acoustics software ODEON[1]. It provides simulations of various detailed 3D room models with parameters (e.g. absorption coefficient, scattering coefficient) measured in the real world. We selected two rooms: the PTB music studio of which the volume is approximately 400 $m^3$ (8 $m$ * 10 $m$ * 5 $m$), and the Auditorium21 at Technical University of Denmark (DTU) of which the volume is approximately 1200 $m^3$ (15 $m$ * 12 $m$ * 7 $m$). These two rooms are representative and have been used in several acoustics projects [5, 6], and they fit well the potential application environments where our method can be used to create adequate room acoustic effects.

To gather a collection of high-quality IRs, we simulated 1000 IRs with 25 omni-directional sources and 40 receivers in each room. We set the source-receiver distances and the receiver-surface distances according to ISO 3382 [1]. We simulated the IRs without background noise at a temperature of 20 °C and a relative humidity of 50%. The other simulation parameters were set as default in ODEON[2].

In order to investigate the relationships $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$ after the IR simulation, we applied an octave filter bank to filter the IRs into each octave band and we applied a high-pass filter to obtain the ultra-IR (20kHz - 22kHz). We determined the slopes of the IRs by fitting a straight line to the IR waveforms in the dB logarithmic scale, and we determined the RT60 values using the well-known Schroeder method [22].

We calculated the ratio $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$ based on the set of 1000 simulated IRs in each room. In Table 1 and Table 2, we summarize the mean (M) and standard deviation (SD) of these two ratios, and calculate the $f_{slope}$ using Eq(1). The mean values of $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$ vary significantly across octave bands, but the standard deviation in each octave band is small. This result indicates the rationality to take $RT60_{ultra}/RT60_{octave}$ as an *octave constant* for each octave band. Additionally, $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$ are inversely related, so the resultant $f_{slope}$ values are similar in each octave band, as shown in Table 1 and Table 2. Therefore, we computed an average $f_{slope} = 0.939$ and an average $f_{slope} = 0.915$ as the *room constant* for PTB music studio and for Auditorium21.

*3.2.2 Determining the Slope Factor and the RT60 Ratios in Real Rooms.* In Section 3.2.1, we demonstrate the rationality of determining $f_{slope}$ and $RT60_{ultra}/RT60_{octave}$ with a large amount of IRs in simulated rooms. In this section, we explore the feasibility of this approach in real rooms with a small number of IRs (16 in our case) for practical reasons. We chose two rooms in our institution buildings as shown in Figure 4: a classroom of approximately 220 $m^3$ and a lounge of approximately 125 $m^3$. In each room, we placed a FOCAL Shape 65 loudspeaker[3] at one corner with source-surface distances of approximately 1 $m$. The loudspeaker was put to face the center of the room. We recorded 16 IRs using a Primo EM172 microphone module[4] at 16 arbitrarily chosen locations with the consideration of source-receiver distance and the receiver-surface distance as recommended in [1].

Like in the case of simulations, we calculated $RT60_{ultra}/RT60_{octave}$, $S_{ultra}/S_{octave}$, and $f_{slope}$. Results are shown in Table 3 and Table 4. As expected, the deviations from the average values are slightly larger compared to the results from the simulations. Considering the significantly smaller number of measured IRs, more complex sound propagation and interaction processes in the real world, and

**Table 1: PTB Music Studio: The mean (M) and standard deviation (SD) of $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$ across 1000 IRs. The $f_{slope}$ is calculated using Eq(1) based on the mean values of $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$.**

| | U/63 | U/125 | U/250 | U/500 | U/1k | U/2k | U/4k | U/8k | U/16k |
|---|---|---|---|---|---|---|---|---|---|
| RT60 M | 0.558 | 0.504 | 0.431 | 0.447 | 0.486 | 0.463 | 0.533 | 0.707 | 0.956 |
| RT60 SD | 0.046 | 0.028 | 0.017 | 0.013 | 0.014 | 0.011 | 0.011 | 0.014 | 0.015 |
| slope M | 1.958 | 2.031 | 2.232 | 2.093 | 1.875 | 1.887 | 1.608 | 1.235 | 0.957 |
| slope SD | 0.160 | 0.148 | 0.146 | 0.111 | 0.079 | 0.061 | 0.040 | 0.028 | 0.018 |
| $f_{slope}$ | 1.092 | 1.024 | 0.962 | 0.937 | 0.911 | 0.875 | 0.857 | 0.873 | 0.916 |

**Table 2: Auditorium21: The mean (M) and standard deviation (SD) of $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$ across 1000 IRs. The $f_{slope}$ is calculated using Eq(1) based on the mean values of $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$.**

| | U/63 | U/125 | U/250 | U/500 | U/1k | U/2k | U/4k | U/8k | U/16k |
|---|---|---|---|---|---|---|---|---|---|
| RT60 M | 0.708 | 0.696 | 0.463 | 0.344 | 0.311 | 0.334 | 0.434 | 0.642 | 0.938 |
| RT60 SD | 0.062 | 0.046 | 0.020 | 0.010 | 0.007 | 0.007 | 0.009 | 0.014 | 0.015 |
| slope M | 1.553 | 1.504 | 2.046 | 2.548 | 2.704 | 2.446 | 1.886 | 1.331 | 0.983 |
| slope SD | 0.120 | 0.102 | 0.125 | 0.145 | 0.124 | 0.082 | 0.050 | 0.029 | 0.015 |
| $f_{slope}$ | 1.099 | 1.048 | 0.948 | 0.878 | 0.843 | 0.819 | 0.820 | 0.855 | 0.923 |

(a) classroom (about 10m*6.4m*3.4m) (b) lounge (about 6m*6m*3.4m)

**Figure 4: Two real rooms used in our study.**

[1] https://odeon.dk/

[2] https://odeon.dk/download/Version15/OdeonManual.pdf

[3] https://www.focal.com/en/pro-audio/monitoring-speakers/shape/monitoring-speakers/shape-65

[4] https://micbooster.com/audio-cable/127-primo-em172-with-35-mm-plug.html?search_query=primo+EM172+3.5mm+plug&results=60

**Table 3: Real classroom: The mean (M) and standard deviation (SD) of $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$ across 16 IRs. The $f_{slope}$ is calculated using Eq(1) based on the mean values of $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$.**

| | U/63 | U/125 | U/250 | U/500 | U/1k | U/2k | U/4k | U/8k | U/16k |
|---|---|---|---|---|---|---|---|---|---|
| RT60 M | 0.257 | 0.313 | 0.282 | 0.275 | 0.272 | 0.271 | 0.276 | 0.376 | 0.689 |
| RT60 SD | 0.023 | 0.026 | 0.019 | 0.018 | 0.019 | 0.018 | 0.018 | 0.024 | 0.028 |
| slope M | 2.815 | 2.452 | 2.383 | 2.742 | 2.387 | 2.224 | 2.038 | 1.498 | 0.956 |
| slope SD | 0.219 | 0.215 | 0.121 | 0.193 | 0.132 | 0.123 | 0.124 | 0.090 | 0.035 |
| $f_{slope}$ | 0.726 | 0.769 | 0.673 | 0.756 | 0.650 | 0.604 | 0.563 | 0.564 | 0.659 |

**Table 4: Real lounge: The mean (M) and standard deviation (SD) of $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$ across 16 IRs. The $f_{slope}$ is calculated using Eq(1) based on the mean values of $RT60_{ultra}/RT60_{octave}$ and $S_{ultra}/S_{octave}$.**

| | U/63 | U/125 | U/250 | U/500 | U/1k | U/2k | U/4k | U/8k | U/16k |
|---|---|---|---|---|---|---|---|---|---|
| RT60 M | 0.231 | 0.286 | 0.296 | 0.303 | 0.290 | 0.284 | 0.324 | 0.451 | 0.753 |
| RT60 SD | 0.018 | 0.027 | 0.023 | 0.024 | 0.019 | 0.019 | 0.021 | 0.031 | 0.037 |
| slope M | 2.824 | 2.431 | 2.180 | 2.085 | 1.964 | 1.838 | 1.652 | 1.180 | 0.804 |
| slope SD | 0.275 | 0.247 | 0.177 | 0.162 | 0.151 | 0.146 | 0.142 | 0.097 | 0.046 |
| $f_{slope}$ | 0.653 | 0.697 | 0.646 | 0.632 | 0.571 | 0.522 | 0.535 | 0.533 | 0.606 |

the higher noise level than in the simulation, we argue that the results demonstrate the feasibility of using $RT60_{ultra}/RT60_{octave}$ as an *octave constant* and the feasibility of determining $f_{slope}$ as a *room constant* for real rooms. We computed an average $f_{slope} = 0.663$ for the classroom and $f_{slope} = 0.600$ for the lounge.

In summary, we proposed and demonstrated the feasibility of calculating two parameters, the *room constant* $f_{slope}$ and the *octave constants* $RT60_{ultra}/RT60_{octave}$, from a small number of IRs that are pre-measured in an environment. These parameters can then be used to calculate $S_{octave}$ using Eq(1) to determine the exponential function for octave-IR initialization. For AR applications, these room-dependent parameters can be stored online e.g. in the Open AR Cloud[5], and retrieved when a user enters the space for the synthesis of room IRs at arbitrary locations in the space.

## 3.3 Correction of Early Reflections

Similar to the parametric wave field coding method [17], we take 200$ms$ after the window of the direct sound as the duration of the early reflections (ER). The remaining section that follows the 200$ms$ period is considered the late reverberation (LR).

After the initialization, we first calculate the total energy of each initialized octave-IR as the target energy of the finalized IR in each octave band. Then, to determine the target energy of ER and LR, we explored the energy ratio $energy_{ER}/energy_{LR}$ in the simulated and the real rooms. As before, we calculated the relationship of the energy ratio between the ultra-IR and each octave-IR $energy - ratio_{ultra}/energy - ratio_{octave}$. We found that in all four rooms, the value $energy-ratio_{ultra}/energy-ratio_{octave}$ is approximately 1 with a small deviation for all octave bands: (1) PTB music studio ($M = 1.057, SD = 0.030$); (2) Auditorium21 ($M = 1.098, SD = 0.088$); (3) real classroom ($M = 1.034, SD = 0.024$); (4) real lounge ($M = 1.009, SD = 0.006$). Therefore, given a new input ultra-IR, we will calculate its $energy_{ER}/energy_{LR}$, and use this ratio together with the target total energy of the octave-IR to adjust the ER and LR.

As in [17], we regard the ER as a sum of specular and diffuse components, and the total energy of these two components is supposed to match the target ER energy. We take the initialized GWN in the ER part as the diffuse component and scale the sample amplitudes to take only 10% of the total ER energy. Then, the specular component will take 90% of the total ER energy.

Ideally, one could detect the occurrences and the amplitudes of strong reflections in the ultra-IR and construct the specular components in each octave band accordingly. However, given the difficulty to accurately capture the reflection occurrences from an ultra-IR due to the strong attenuation and absorption in the ultra-frequency band, we use the technique described in [17]: generating a set of 250 peaks with prime number sample delays. We generate 250 prime number sample delays within the 200$ms$ period, and use the same set of 250 delays for each octave band to create the specular component. Different from Raghuvanshi and Snyder's implementation [17] of assigning random amplitudes to these specular samples, we first assign random amplitudes within the range $[-1, 1]$ and then apply the same exponential function as used for initialization in order to maintain the envelope of the octave-IR. Finally, we scale the specular samples to take 90% of the total ER energy.

After the above steps, the ER part of each octave-IR can be considered "corrected".

## 3.4 Adjustment of Late Reverberation

The LR part of an IR can be taken as a pure diffuse process and can be generated using a GWN with an exponentially decaying envelope [25]. Our initialization has created such a LR for each octave-IR. In this step, we scale the sample amplitudes in the LR to meet the energy requirement as discussed in the previous section.

**Summary:** After the above four steps, each octave-IR is normalized, summed, and finally normalized as the synthesized room IR for the given source-receiver position. This IR synthesis method has potential to be used for real-time interactive audio AR applications. First, as discussed before, the pre-calculated parameters can be stored online and retrieved at runtime to synthesize IRs at arbitrary user locations in an environment. Secondly, synthesizing a room IR only takes around 100$ms$ with an un-optimized Python implementation on a laptop (Intel Core i7-6700 2.6GHz CPU, 8G RAM). This indicates the possibility of real-time IR synthesis with low-level optimized code, which has potential even for wearable implementation. Moreover, there is potential to capture ultra-IRs using smartphones. The impulses can be played from an ultrasonic loudspeaker that is physically placed at the location for the virtual sound source in the AR application. The method is illustrated with audio samples in the accompanying video[6].

# 4 OBJECTIVE EVALUATION

To evaluate the proposed method, we compared the synthesized IRs with the ground truth IRs with regard to four common [9, 20] acoustic parameters: RT60, early decay time (EDT), clarity 80 ($C_{80}$), and center time ($T_S$). RT60 describes how a room would overall

[5]https://www.openarcloud.org/

[6]https://youtu.be/aOlUEW23T_A

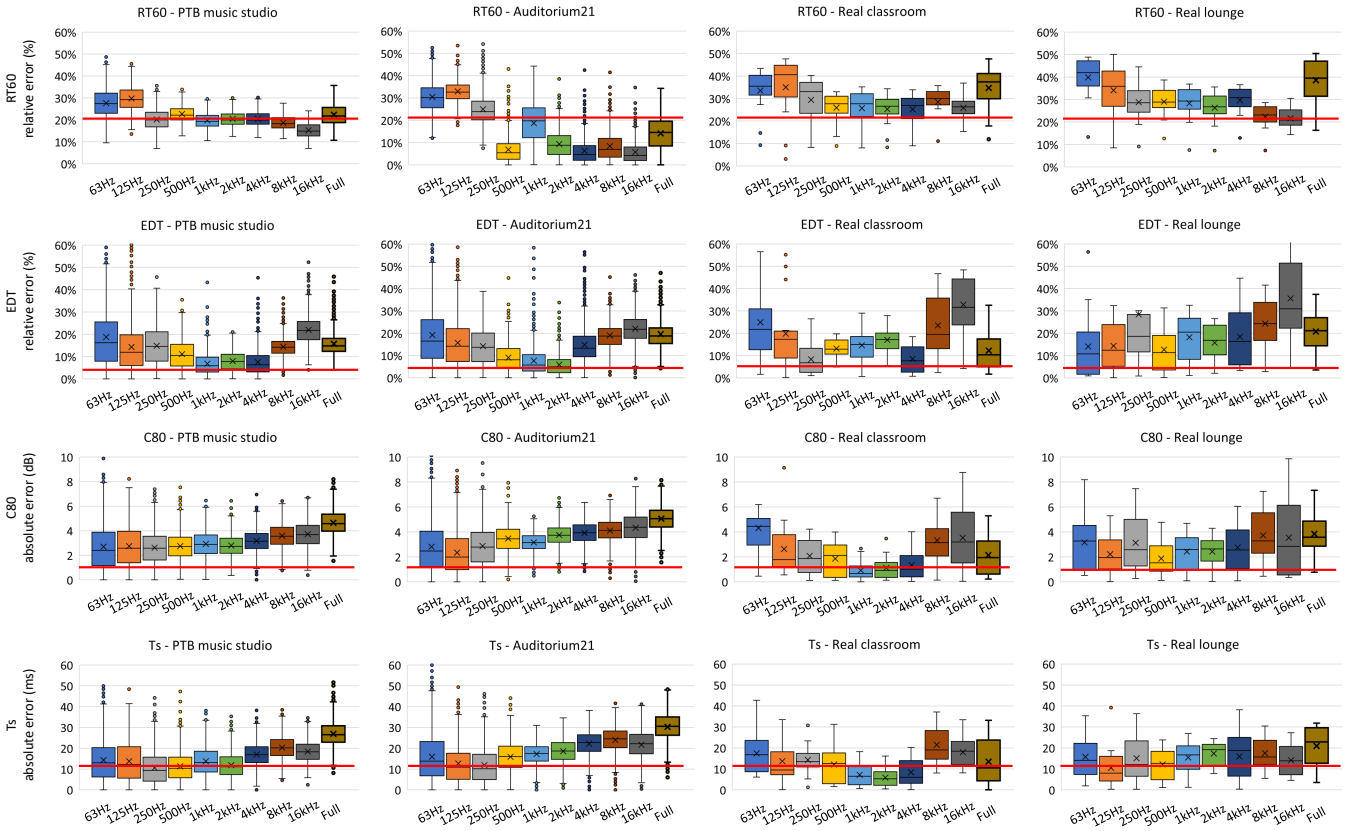


**Figure 5: Differences between the synthesized IRs and the ground truth IRs with respect to the acoustic parameters RT60, EDT, $C_{80}$, and $T_s$ in the simulated rooms and the real rooms. In each plot, we show the difference for each octave-IR and for the complete room IR (label: Full). According to literature, the JNDs were chosen to be relative 20% for RT60, relative 5% for EDT, absolute $1dB$ for $C_{80}$, and absolute $10ms$ for $T_S$. In each plot, we highlight the JND value with a red line.**

"color" a sound with its size, shape, and absorption. EDT relates to the energy in the early reflections. $C_{80}$ evaluates the clarity level of a room. $T_S$ relates to the balance between clarity and reverberance.

Since human auditory perception is only sensitive to a certain level of difference between two sounds, we compare the difference between the synthesized IR and the ground truth IR according to the just noticeable difference (JND) values. In literature, there exist several JND standards as measurements were conducted under different conditions and for a range of parameters. In this evaluation, we chose a relatively strict standard considering the acoustic environment in the study and potential applications of the IR synthesis approach: 20% for RT60 [16], 5% for EDT [1], $1dB$ for $C_{80}$ [1], and $10ms$ for $T_S$ [1].

We conducted the objective evaluation with 1000 newly simulated IRs in PTB music studio and Auditorium21, and with 16 newly measured IRs in the real classroom and lounge as the ground truth IRs, from which we extracted the octave-IRs and ultra-IRs. Given the extracted ultra-IRs, we used the parameters $f_{slope}$ and $RT60_{ultra}/RT60_{octave}$ as calculated before to synthesize IRs following the approach as described in Section 3. We calculated the difference between each ground truth IR and the corresponding synthesized IR in terms of the above four parameters. Figure 5 demonstrates all the differences in box plots.

The average RT60 values in each room are $1.02s$ (PTB music studio), $1.02s$ (Auditorium21), $0.64s$ (real classroom), and $0.63s$ (real lounge). Compared with the real rooms, simulated rooms in general show a slightly more concentrated error distribution and most errors are around or even smaller than the JND. For all four rooms, the RT60 of most synthesized IRs are $0.1s - 0.3s$ shorter than the RT60 of the ground truth IRs. This could be due to the insufficient energy in the LR phase. As discussed in Section 3, with the $200ms$ ER duration, the energy ratio between the ER and the LR was very similar in different frequency bands. However, lower frequencies tend to have a marginally higher proportion of LR energy. Correspondingly, as shown in Figure 5, errors in lower frequency bands are generally larger. A shorter RT60 might make a listener perceive a smaller environment size than it is supposed to be.

Synthesized IRs generally have a longer EDT than the ground truth IRs, which means that the synthesized IRs have a slower energy decay until $-10dB$ in the Schroeder's backward integral compared with the ground truth IRs. For a continuous sound source (e.g. a piece of music), a longer environmental EDT indicates that a

sound incident might mask over the next sound incidents from the source. This could influence the perceived clarity of the sound.

Most synthesized IRs have a smaller $C_{80}$ value than the ground truth IRs. This indicates that compared with the ground truth IRs, synthesized IRs have less relative energy before 80ms. Considering that our method takes 200ms as the ER duration and generates specular components at prime number delays within this 200ms time window, the IR synthesis process might produce more energy between 80ms and 200ms than it is supposed to be. According to the definition of $C_{80}$ [1], a sound blurring effect will be stronger with more reflections later than 80ms (i.e. smaller $C_{80}$ value). Therefore, we assume that the synthesized IRs will render sounds with a marginally lower clarity.

$T_S$ is also a clarity parameter that describes the balance between clarity and reverberation of a room. Small $T_S$ values indicate a clear room while large $T_S$ values indicate a reverberant room. Compared with the ground truth IRs, most synthesized IRs have a larger $T_S$ value. This indicates that the synthesized IRs might create a more reverberant sound effect, which corresponds with the results of EDT and $C_{80}$ as discussed above.

So far, we have discussed the objective evaluation results in terms of four acoustic parameters with their corresponding JND values. We have also discussed possible reasons and potential effects of the results. Since JND standards were obtained from previous research under specific measurement conditions, the objective evaluation results might not completely reflect a user's auditory perception of the sounds rendered with our synthesized IR. To evaluate a user's auditory experience, we present a user study in next section.

## 5 USER EVALUATION

We conducted an online user study to compare the synthesized IRs with the ground truth IRs. In the following, we describe the study design and discuss the study results.

### 5.1 Study Design

In each simulated and real room, we selected one synthesized IR of which the results of objective evaluation were approximately at the average level among all IRs in the room. We selected four dry sounds to be augmented with the IRs – two music and two voice samples: symphony [7] (duration: 28*s*), guitar solo[8] (duration: 9*s*), female voice[9] (duration: 10*s*), and male voice[10] (duration: 10*s*). Both the symphony and the guitar music have a wide frequency spectrum. The female and the male voice delivers the same content. For each dry sound, we implemented convolutions with the synthesized IRs and with their corresponding ground truth IRs (simulated/recorded). Therefore, there was a total of 16 pairs of comparison, in which the two sounds were labeled as Test1 and Test2. To avoid bias, we shuffled the order of Test1 and Test2 in each pair, and the room number was not shown to the participants. The dry version of each sound was provided as a reference. For each pair of comparison, participants were allowed to listen to the sounds as many times as needed, and answered the following questions:

[7] https://users.aalto.fi/~ktlokki/Sinfrec/sinfrec.html
[8] http://www.lam.jussieu.fr/Projets/index.php?page=AVAD-VR
[9] http://www-mmsp.ece.mcgill.ca/Documents/Data/
[10] http://www-mmsp.ece.mcgill.ca/Documents/Data/

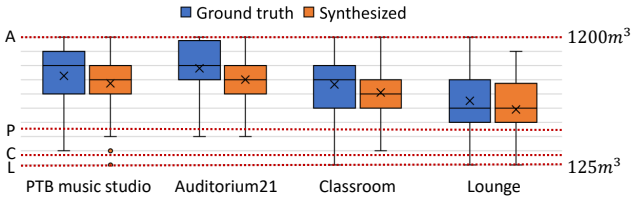


**Figure 6: Participants' estimation of the room size in the range $[125m^3, 1200m^3]$. Four red dash lines indicate the ground truth size of the PTB music studio (P), Auditorium21 (A), classroom (C), and lounge (L).**

(1) *I can easily recognize the difference between the dry sound and Test1 (Test2).* Participants answered these two questions on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree).

(2) *For Test1 (Test2), I feel the sound is played in a room of the following size.* We provided a 10-point linear scale from the smallest size ($6\,m * 6\,m * 3.4\,m$, the real lounge) to the largest size ($15\,m * 12\,m * 7\,m$, the simulated Auditorium21) of the four rooms.

(3) *Compare Test 1 and Test 2, how similar are they?* Participants answered this question on a 5-point Likert scale from 1 (very different) to 5 (very similar).

The first two questions aimed to confirm that participants could perceive the acoustic effects in the test sounds. The room size might be difficult to estimate, but participants might perceive the room size of Test1 and Test2 differently, according to the objective evaluation results. The last question asked participants' perceptual similarity between the synthesized and the ground truth sound.

### 5.2 Results and Discussion

A total of 22 participants (7 female, 15 male, age $\in [19, 37]$, $M = 27.7$, $SD = 5.04$) took part in the user study. Only one participant had some acoustics knowledge, while all the others had no background in acoustics or audio/music processing.

As for the first two questions, participants could easily recognize the difference between the dry sound and the ground truth sound ($M = 4.698$, $SD = 0.671$), or between the dry sound and the synthesized sound ($M = 4.659$, $SD = 0.718$).

Figure 6 shows participants' estimation of the room size. The ground truth room sizes are $400\,m^3$ (PTB music studio), $1200\,m^3$ (Auditorium21), $220\,m^3$ (classroom), and $125\,m^3$ (lounge). Figure 6 shows a big deviation from the participants' estimation to the correct size of each room. This could be because of two reasons. First, according to participants' feedback, this question was difficult since they were not provided with auditory references for each scale to calibrate their hearing. Secondly, the acoustic properties of the materials in a room might cause an impression of the volume that did not match the real size. We applied an Align Rank Transform (ART) [26] before using a two-way repeated measures ANOVA. The ANOVA test shows that participants' estimation of the room size significantly changed across rooms ($p < 0.001$), which indicates that participants recognized some size differences in different rooms. The ANOVA test also shows that participants' estimation based on the synthesized IRs was significantly smaller than the corresponding estimation based on the ground truth IRs ($p = 0.001$).

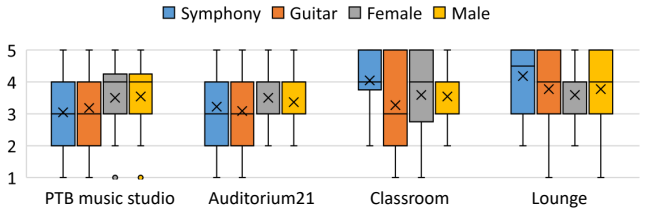


**Figure 7: Participants' perceptual similarity between the ground truth and the synthesized sounds in each room for each audio sample (1: very different, 5: very similar).**

As discussed in Section 4, this could be because the synthesized IRs generally had a shorter reverberation time, which could make participants perceive a smaller space size.

As shown in Figure 7, participants overall leaned towards a similar perception between the ground truth and the synthesized IRs since all cases show an average similarity score over the neutral scale 3. For each room, we conducted a Friedman test and found that participants' perceptual similarity did not vary significantly across the four sound samples (all $p > 0.152$). According to the participants, they generally judged the similarity based on their perception of bass, clarity, and reverberance in the rendered sounds. Note that the perception of similarity was subjective, and this evaluation might be influenced by the audio interface, the headphone, and the environment where the participant did the study. These could be the reasons for the large deviations across individuals, but the overall result shows a positive trend towards similar perception. This indicates that our method has the potential to produce perceptually acceptable room acoustics. According to the previous research [3], providing such a perceptually adequate representation of the acoustic environment is likely to suffice in AR applications, where a user's visual sense works in concert with the auditory sense to perceive a complete picture of the environment.

## 6 CONCLUSION AND FUTURE WORK

We proposed a fast parametric method to synthesize full-frequency IRs from ultrasonic measurements to render perceptually adequate acoustic effects for AR applications. Instead of modeling a complete sound propagation process, our method approximates IRs using measured ultra-IRs and a few parameters that can be determined from pre-recorded IRs in the environment of interest. We evaluated the synthesized IRs according to common acoustic parameters and demonstrated the efficacy of this method in creating an adequate acoustic environment by conducting a user study. We also discussed the potential of using this method in interactive AR applications as a user moves around in the space, since the IR synthesis runs fast with pre-computed parameters that can be retrieved from online storage at runtime, and the ultra-IRs can be measured live.

This work initiates several interesting directions for future exploration. First, this work shows the potential to create convincing acoustic effects by measuring only ultra-IRs. We intend to further explore how the movement of the user and objects influence the IRs, and we also plan to implement this method on wearable devices in order to test its applicability in real-time AR applications.

## ACKNOWLEDGMENTS

We thank all our study participants for their time and feedback.

## REFERENCES

[1] ISO 3382-1. 2009. Acoustics—Measurement of room acoustic parameters—Part 1: Performance spaces.
[2] Jont B Allen and David A Berkley. 1979. Image Method for Efficiently Simulating Small-Room Acoustics. *The Journal of the Acoustical Society of America* 65, 4 (1979), 943–950.
[3] Will Bailey and Bruno Fazenda. 2018. The Effect of Visual Cues and Binaural Rendering Method on Plausibility in Virtual Environments. In *Audio Engineering Society Convention 144*. Audio Engineering Society.
[4] Vincent Becker, Linus Fessler, and Gábor Sörös. 2019. GestEar: Combining Audio and Motion Sensing for Gesture Recognition on Smartwatches. In *ACM ISWC*. London, UK.
[5] Ingolf Bork. 2005. Report on the 3rd Round Robin on Room Acoustical Computer Simulation–Part I: Measurements. *Acta Acustica united with Acustica* 91, 4 (2005), 740–752.
[6] Claus Lynge Christensen, George Koutsouris, and Jens Holger Rindel. 2013. The ISO 3382 Parameters: Can We Simulate Them? Can We Measure Them?. In *ISRA*. 9–11.
[7] Aki Harma, Julia Jakka, Miikka Tikander, Matti Karjalainen, Tapio Lokki, and Heli Nironen. 2003. Techniques and Applications of Wearable Augmented Reality Audio. In *Audio Engineering Society Convention 114*. Audio Engineering Society.
[8] Vedad Hulusic, Carlo Harvey, Kurt Debattista, Nicolas Tsingos, Steve Walker, David Howard, and Alan Chalmers. 2012. Acoustic Rendering and Auditory–Visual Cross-Modal Perception and Interaction. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 102–131.
[9] Hansung Kim, Luca Hernaggi, Philip JB Jackson, and Adrian Hilton. 2019. Immersive Spatial Audio Reproduction for VR/AR using Room Acoustic Modelling from 360 Images. In *IEEE VR*. 120–126.
[10] Hansung Kim, Richard J Hughes, Luca Remaggi, Philip JB Jackson, Adrian Hilton, Trevor J Cox, and Ben Shirley. 2017. Acoustic Room Modelling using A Spherical Camera for Reverberant Spatial Audio Objects. In *Audio Engineering Society Convention 142*. Audio Engineering Society.
[11] S Kopuz and N Lalor. 1995. Analysis of Interior Acoustic Fields using the Finite Element Method and the Boundary Element Method. *Applied Acoustics* 45, 3 (1995), 193–210.
[12] H Kuttruff. 2000. Room Acoustics, UK.
[13] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-play acoustic activity recognition. In *ACM UIST*. 213–224.
[14] Pontus Larsson, Aleksander Väljamäe, Daniel Västfjäll, Ana Tajadura-Jiménez, and Mendel Kleiner. 2010. Auditory-Induced Presence in Mixed Reality Environments and Related Technology. In *The Engineering of Mixed Reality Systems*. Springer, 143–163.
[15] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. 2018. Scene-Aware Audio for 360 Videos. *ACM TOG* 37, 4 (2018), 1–12.
[16] Zihou Meng, Fengjie Zhao, and Mu He. 2006. The Just Noticeable Difference of Noise Length and Reverberation Perception. In *IEEE ISCIT*. IEEE, 418–421.
[17] Nikunj Raghuvanshi and John Snyder. 2014. Parametric Wave Field Coding for Precomputed Sound Propagation. *ACM TOG* 33, 4 (2014), 1–11.
[18] Nikunj Raghuvanshi and John Snyder. 2018. Parametric Directional Coding for Precomputed Sound Propagation. *ACM TOG* 37, 4 (2018), 1–14.
[19] Lauri Savioja, Dinesh Manocha, and M Lin. 2010. Use of GPUs in Room Acoustic Modeling and Auralization. In *ISRA*. 3.
[20] Carl Schissler, Christian Loftin, and Dinesh Manocha. 2017. Acoustic Classification and Optimization for Multi-Modal Rendering of Real-World Scenes. *IEEE TVCG* 24, 3 (2017), 1246–1259.
[21] Carl Schissler and Dinesh Manocha. 2016. Interactive Sound Propagation and Rendering for Large Multi-Source Scenes. *ACM TOG* 36, 4 (2016), 1.
[22] Manfred R Schroeder. 1965. New method of measuring reverberation time. *The Journal of the Acoustical Society of America* 37, 6 (1965), 1187–1188.
[23] Zhenyu Tang, Nicholas J Bryan, Dingzeyu Li, Timothy R Langlois, and Dinesh Manocha. 2020. Scene-Aware Audio Rendering via Deep Acoustic Analysis. *IEEE TVCG* 26, 5 (2020), 1991–2001.
[24] Nicolas Tsingos, Wenyu Jiang, and Ian Williams. 2011. Using Programmable Graphics Hardware for Acoustics and Audio Rendering. *Journal of the Audio Engineering Society* 59, 9 (2011), 628–646.
[25] Vesa Valimaki, Julian D Parker, Lauri Savioja, Julius O Smith, and Jonathan S Abel. 2012. Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 5 (2012), 1421–1448.
[26] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses using Only ANOVA Procedures. In *ACM CHI*. 143–146.
[27] Jing Yang, Yves Frank, and Gábor Sörös. 2019. Hearing Is Believing: Synthesizing Spatial Audio from Everyday Objects to Users. In *ACM AH*. 1–9.