Minimally Simple Binaural Room Modelling Using a Single Feedback Delay Network

NATALIE AGUS[1], HANS ANDERSON[1], JER-MING CHEN[1], SIMON LUI[1], AND DORIEN HERREMANS[1,2]
(natalie_agus@mymail.sutd.edu.sg) (hans_anderson@mymail.sutd.edu.sg) (jerming_chen@sutd.edu.sg)
(nomislui@gmail.com)(dorien_herremans@sutd.edu.sg)

[1]*Singapore University of Technology and Design, Singapore*
[2]*Agency for Science, Technology and Research (A*STAR), Singapore*

The most efficient binaural acoustic modeling systems use a multi-tap delay to generate accurately modeled early reflections, combined with a feedback delay network that produces generic late reverberation. We present a method of binaural acoustic simulation that uses one feedback delay network to simultaneously model both first-order reflections and late reverberation. The advantages are simplicity and efficiency. We compare the proposed method against the existing method of modeling binaural early reflections using a multi-tap delay line. Measurements of ISO standard evaluators including interaural correlation coefficient, decay time, clarity, definition, and center time, indicate that the proposed method achieves comparable level of accuracy as less-efficient existing methods. This method is implemented as an iOS application, and is able to auralize input signal directly without convolution and update in real time.

0 INTRODUCTION

The widespread adoption of acoustic modeling in contexts such as 3D gaming and virtual reality simulation is hindered by the complexity of the implementation. There exists a great variety of methods for acoustic modeling of virtual spaces, ranging from computationally intensive and very accurate to efficient rough approximations. The goal of the method we present here is to improve on the efficiency and simplicity of the most efficient methods with minimal loss of accuracy.

The most accurate binaural reproduction of the acoustics of a real room is obtained by convolution of a dry input signal with the recorded binaural room impulse response (BRIR). Obviously this method is limited to rooms that exist physically, and of which we can actually record the BRIR. The recorded BRIR depends on listener and source positions, as well as the room shape and placement of objects and materials. It is not possible to record and store BRIRs for all possible combinations of these parameters. Because of these limitations, acoustic modeling is an attractive alternative.

Most acoustic modeling methods fall under one of two categories of algorithms, Numerical Acoustics (NA) and Geometrical Acoustics (GA).

Numerical acoustics comprises various analytical approaches to solving the wave equation. The main benefit of NA methods is that they can account for wave phenomena such as interference and diffraction. However, because they are computationally intensive, it is not yet possible to solve the wave equation for the entire duration of the RIR across all audible frequency bands [1].

Unlike numerical acoustics, geometric acoustics based approaches assume that sound waves propagate as rays. Many of these techniques are adapted from the fields of optics and computer graphics. One of the most widely used geometrical acoustics methods is the Image Source Method (ISM) [2], where, upon contact with a flat surface, we assume that the reflection of sound waves is perfectly specular. Traditional GA methods alone are known to be unable to model the diffraction phenomena that are more prominent in the lower frequency bands where the wavelength of sound exceeds the dimensions of large objects in the room [3]. However, GA methods are able to simulate many other important perceptual qualities and are often more efficient than NA methods. It is possible to combine GA and NA methods together, using the more accurate NA model at low frequencies where complex wave effects are prominent and the GA model in the higher frequency ranges. A typical strategy is to apply a NA method to model the acoustics below the Schroeder frequency, which is around 50Hz for a typical concert hall. Above that frequency, modes of resonance become so dense that it is more appropriate to model them as stochastic processes using GA methods [4].

Applications of both categories of reverberation algorithms include acoustic simulation for training simulations,

Preprint of Angus, N, Anderson, H, Chen, JM, Lui, S, Herremans, D. 2018. Perceptual evaluation of measures of spectral variance. Journal of the Audio Engineering Society. In Press.

music recordings and computer games. Since there is a trade-off between accuracy and computational complexity, the appropriate choice of method for simulating room acoustics depends on the specific requirements of each application.

The method we propose here falls under the category of GA methods. It combines the Acoustic Rendering Equation (ARE) [5] and a Feedback Delay Network (FDN) [6] with a bank of head related transfer function filters (HRTF) [7] and a bank of interaural time difference (ITD) delay lines to simulate binaural room acoustics in real time.

1 RELATED WORK

Many acoustic modeling systems work by precomputing impulse responses offline and caching them. This paper focuses on methods that are efficient enough to update in realtime without caching a database of precomputed impulse responses.

Several GA approaches allow modeling parameters to update in real time by combining a detailed early reflections with a generic reverb structure that produces diffuse late reverberation. In those cases the late reverb is produced either by convolution or by an efficient algorithmic reverberator. The most widely used algorithmic reverberators are feedback delay networks (FDN), which are efficient and produce good quality sound output [1, 8].

This category of hybrid GA approaches includes methods that range from simple auralization algorithms to extensive room modeling systems such as DIVA [9, 10] and RAVEN [11, 12]. These auralization programs enable users to navigate in real time through a virtual environment.

The DIVA auralization system utilizes a mixture of offline and online algorithms [9]. The system is modularised into an ISM-based early reflection unit that is frequently updated based on user input and location, and a late reverb unit that uses an FDN-like structure with precomputed coefficients based on room acoustical parameters to produce late reverb impulse responses. These coefficients are obtained from the combination of a numerical finite difference method applied to low frequencies and geometrical ray tracing method applied to high frequencies. They account for air absorption and acoustic properties of various materials.

The rationale for using a generic late reverb unit without emphasis on detailed individual reflections is that the late reverb is thought to contain diffuse, random reflections, with an exponentially decaying envelope [13]. Since human listeners can not perceive the detail of individual reflected rays in such a complex acoustic phenomenon, it is difficult for them to perceive any difference between a detailed model and a generic approximation of late reverb. Separate delay lines for interaural delay and minimum phase head-related transfer function filters are used to reproduce binaural effects, whose coefficients are obtained from a database keyed according to azimuth and elevation, derived using measurements from human subjects.

RAVEN differs from DIVA in the way it produces the late reverb using stochastic geometrical modeling methods to generate an impulse response, instead of using an FDN [11]. Stochastic ray tracing is used to compute the time-energy profile of the late reflections, which is then used to generate filters which, when applied to a noise signal, produce a reverb impulse response. In stochastic ray tracing, a random decision between pure specular reflection or diffuse reflection towards a random direction is taken each time a ray encounters a surface [14]. This method prevents the number of rays in the simulation from growing exponentially in the length of the impulse response. For early reflections, RAVEN also uses the image source method, accelerated using binary space partitioning (BSP), that allows fast visibility checks of the image sources, and therefore enabling real-time updates [15]. RAVEN updates its early reverb simulation more frequently than the late reverb.

In [16], Menzer introduced a real-time binaural room simulation algorithm that is efficient enough to run on mobile devices, and directly processes the input signal without convolution. The work presented in [16] is a less detailed acoustic model than those of DIVA and RAVEN. To enable efficient auralization with minimal computational load, the late reverb does not vary with listener or source position in the room. To do this, Menzer utilized a modified Jot reverberator whose coefficients are obtained from a method of interaural coherence matching using a referenced BRIR [17]. The work in [17] offers an alternative method to compute the coefficients using a single-channel reference RIR and a pair of HRTFs in the case where a stereo reference BRIR is not available. The early reflections are produced using *ISM* up to the second order, followed by convolution with a bank of head-related impulse responses. If the simulation is restricted to perfectly rectangular rooms, the implementation of the *ISM* can be further simplified and the computationally expensive visibility checks can be omitted, allowing for real-time updates.

Menzer proposed another method using two parallel feedback delay networks, one for rendering the early part of the BRIR and the other for the late part [18]. That method is too complex to run on mobile devices at the time the paper was written. The reason for using two FDNs in parallel is that the author observed some diffusion even at the beginning of measured impulse response. The conventional way of connecting the outputs of early reverb units to an FDN results in unrealistically distinct early reflections. This is an especially serious problem when using the image source method because the pure specular reflection model has lower echo density than methods that permit diffusion. The second FDN, used to produce the late reverb, is similar to the one used in his earlier paper [17], but is designed such that it produces higher echo density from the beginning and its parameters do not vary depending on listener and source position. The first FDN produces exact first and second order reflections, modeled by the *ISM*. A small set of head related impulse response convolvers, one pair for each 1^{st} order reflection, produce the binaural signal.

Wendt et. al introduced another computationally efficient and perceptually plausible hybrid binaural room simulation algorithm using *ISM*, FDN, and convolution with *HRIR*s [19]. In this work, the authors modeled the effect of room geometry and wall absorption coefficients in the late reverb, and also incorporate interaural effects in it using HRTFs. This is unique because the late reverb in previous efficient real time simulations does not respond to changes in those parameters and would not respond to 6 degree-of-freedom head movements and rotations like this method does.

To spatialize the late reflections, they use a 12-delay line FDN, where each pair of delay lines corresponds to the length of one of the six major room surfaces, (four walls, ceiling and floor). Due to this arrangement, the method applies only to rectangular room simulations. The output of each channel of the FDN are connected to a series of reflection filters and HRTF filters, before mixing with the outputs of the early reflections unit to form a complete binaural impulse response. The authors present extensive objective and subjective evaluation results. Their method produces good results in terms of Interaural Cross-Correlation Coefficient (IACC_{E3}), however, the authors report a deviation of between 2 to 10 Just Noticeable Differences (JNDs) in terms of Clarity, Definition, and Early Decay Time, measured according to the standards in ISO 3382-1 [20]. The listening test shows that the method has good perceptual accuracy, compared to the measured BRIRs. However this method in [19] are unable to directly auralize the input signal. The time to produce BRIRs of lengths 0.73s and 14.0s for further convolution were 0.71s and 6.80s, respectively.

In [21], Bai et. al proposed a hybrid artificial reverberator called the Acoustic Rendering Network (ARN). It uses the Acoustic Rendering Equation (ARE) and an FDN, and it can theoretically model both specular and diffuse reflections for rooms of arbitrary shape. In contrast to all of the methods mentioned above, Bai models both early reflections and late reflections using a single FDN, rather than using a separate early reflections unit consisting of multi-tap delay lines such as the one presented in [9]. This is done by first discretizing the room surfaces into patches and then separating the reflection paths into three parts: one from the source to each patch, one from patch to patch, and one from each patch to the listener. The ARE is then used to determine the amount of energy received by each patch from the source and other patches, and also the total energy received at the listener position. The feedback matrix is set such that each coefficient corresponds to the amount of energy exchanged between a pair of patches. If N represents the number of patches in the surface geometry model then the Bai et al method requires a mixing matrix of size $2N + N^2$. The authors reported that the method takes 16.5s to synthesize a 1 second RIR in a rectangular room sized $4m \times 6m \times 4m$ that was discretised into 32 square patches.

In this paper we propose a binaural reverberator that supports arbitrary room shapes, does fast real time parameter updates, and is efficient enough to run on mobile devices. In comparison to related methods, similar advantages are achieved by [19, 16, 21, 9] but only the proposed method achieves all of them simultaneously. Our method produces both early reflections and late reverb using a single FDN without using a separate multi-tap delay for early reflections. This idea of compact design is also proposed in [21]. The most significant difference between the method presented here and the one in [21] is described in section 2 where we show how the proposed method allows us to use a standard unitary mixing matrix such as the the Hadamard matrix for the FDN, while still modeling position-dependent interaural effects not only in early reflections but also in late reverb. This allows us to minimize computation time and enables the proposed method to directly process the input signal in real time rather than using convolution with an impulse response.

2 METHOD

2.1 Method Overview

Figure 1 shows a flowchart diagram illustrating the proposed method. The key innovation in this design is that the lengths of the delay lines in the FDN are set using an acoustic model so that the first impulse out of each delay in the network represents one explicitly modeled first-order reflection. Subsequent circulation of the signal around the FDN produces higher order reflections with less accuracy. The gain coefficients at the input and output of each delay ensure that each early reflection has the correct sign and amplitude.

We use the Acoustic Rendering Equation (ARE) to compute the coefficients μ and υ shown in Figure 1. In this way, the first reflections to issue out of the FDN are exactly as modeled by the ARE. Late reflected energy approaches a state of approximately even diffusion [22] and therefore individual late reflections need not modeled in detail. The proposed method models only the first order reflections in detail; for late reverb, we assume that energy is evenly diffused. Based on that assumption, we approximate the average late reflected energy that reaches the listener from each patch of the discretized geometry. The weakness of this approach in relation to related methods is that the second order reflections not modeled accurately. We will show that this sacrifice leads to a much more efficient design that still gives listeners a natural and plausible sense of location in the acoustic space.

We model the energy flux from each surface geometry patch to the listener proportional to the projected area of the patch as seen from the listener position and inversely proportional to the square of the distance.

In the remaining parts of this section we will explain the mathematics we use to model early reflections and estimate late reverb energy flux for each surface geometry patch. The goal is to calculate the gain coefficients at the inputs μ_n and outputs υ_n of the N delay lines in the FDN.

2.2 The Acoustic Rendering Equation

Our model of 1^{st} order reflections is a standard application of the Acoustic Rendering Equation (ARE) [5]. We refer readers to our previous work in [23] for further details on how we model the first order reflections using the ARE. In this paper, we use the same notations as that in our previous work in [23].

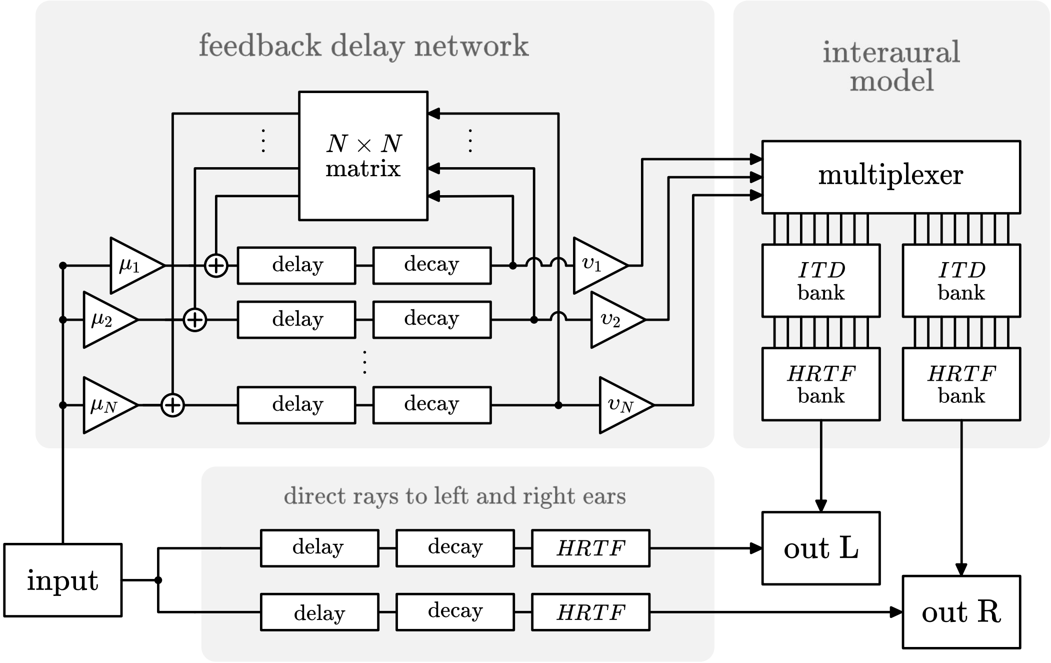

Fig. 1: The proposed system: the delay lengths and output gain coefficients in the FDN are chosen so that the first impulses to come out from the network are the early reflections as modeled by the Acoustic Rendering Equation. Each delay line in the network corresponds to one patch of surface geometry in a 3D model of an acoustic space. By setting appropriate gain coefficients μ_n at the input and υ_n at the output, we simultaneously get a detailed model of first order reflections and an approximated model of late reverb energy energy flux reflecting off each surface.

2.3 Irradiance at the Listener Position from Late Reverb

The following function expresses denotes the acoustic irradiance (energy flux) at the listener position L due to energy reflected off of A_n, the n^{th} surface patch in the 3D model,

$$E(A_n, L) = \left(\frac{N\,\Phi(F_n)}{\pi \mathcal{G}}\right) \int_{A_n} h(\boldsymbol{x}, L)\,\mathrm{d}\boldsymbol{x}. \qquad (1)$$

$\Phi(F_n)$ represents the energy flux output of F_n, the n^{th} channel of the FDN, and $\mathcal{G}$ represents the total area of all surface geometry in the model.

Equation (1) is based on the assumption that as time progresses the reverberated energy increasingly approaches an evenly diffused and mixed state [22]. Therefore average reflected energy flux density of late reverb is assumed to be the same across all surfaces in the 3D room model. Signals in the FDN behave similar to the assumption stated above. If the initial distribution of energy among its N input channels is uneven, after circulating through the mixing matrix, the energy in each channel is approximately the same.[1]. Taking the output of each channel of the FDN to represent the energy flux density at one of the discrete surface patches in our 3D room model and assuming diffuse reflection, we can approximate the acoustic intensity at the listener location that results form the reflected energy coming from each of the surface patches.

Therefore $N\,\Phi(F_n)$ is the combined energy flux output of all N channels of the FDN. Dividing by $\mathcal{G}$, the quantity $N\,\Phi(F_n)/\mathcal{G}$ is the average late reverb energy flux per unit surface area.

The integral in the right hand side of equation (1) represents how much surface area in the room contributes to energy collected at L. The $1/\pi$ term is derived from the conservation of energy of an ideally diffused reflection, where flux input and output at a surface point to all angles is equal if there is zero absorption loss. A full derivation is shown in [23].

$h(\boldsymbol{x}, L)$ is the point collection function, similar to what is defined in our previous work [23], with the addition of the absorption term ξ,

$$h(\boldsymbol{x}, L) = \xi(\boldsymbol{x}, L)\ \mathscr{V}(\boldsymbol{x}, L)\ P(\boldsymbol{x}, L). \qquad (2)$$

The absorption ξ and visibility $\mathscr{V}$ terms are defined as in [5]. The geometry term $P(\boldsymbol{x}, L)$ is also defined as in [23].

The constant N in (1) is the number of discretized surface patches in the 3D model and also the number of channels in the FDN. Because N applies to both the FDN and the discretization of the 3D model, our choice of mixing matrix for the FDN restricts our options for modeling the room. To efficiently achieve maximally even mixing, we use the Fast Hadamard Transform to do the mixing operation, which requires the N be a power of two. Another option which would allow more freedom in the choice of N is

[1] We must select an appropriate mixing matrix to ensure that this is true. One example is the Hadamard matrix [24]

the block-circulant mixing matrix proposed in [25], which requires only that N be a multiple of some integer K, but needs more time to reach an evenly mixed state when K is small.

2.4 Gain Coefficients υ_n at the FDN Output

Let F_n denote the output of the n^{th} channel in the FDN. Since the FDN operates in units of sound pressure, not energy flux, we have the following relation between the energy flux $\Phi(F_n)$ and sound pressure F_n,

$$F_n{}^2 = \Phi(F_n). \tag{3}$$

We define the gain coefficient υ_n as follows,

$$\upsilon_n = \sqrt{\frac{N}{\pi \mathcal{G}} \int_{A_n} h(\boldsymbol{x}, L) \mathrm{d}\boldsymbol{x}}. \tag{4}$$

We can confirm by inspection that the following relation holds,

$$E(A_n, L) = (F_n \upsilon_n)^2. \tag{5}$$

This indicates that multiplying the output of the n^{th} channel of the FDN by υ_n yields the late reverb sound pressure output of the n^{th} surface geometry patch as perceived at the listener position, L.

2.5 Irradiance at the Listener Position from Early Reflections

We first need to discretise the surface geometry $\mathcal{G}$ into a set of a total of N discrete patches $A_n \subset \mathcal{G}$, for $n = 1...N$ and model the 1^{st} order reflection using the ARE. Afterwards, we need to collect that energy at the listener position.

In equation (6) below, $E_n(A_n, L)$ denotes the *acoustic irradiance* at the listener position L due to 1^{st} order emitted radiance ℓ_1 at A_n, the n^{th} surface patch in our 3D model. Irradiance is a measure of incident energy flux per unit area,

$$E_1(A_n, L) = \int_{A_n} h(\boldsymbol{x}, L) \ell_1(\boldsymbol{x}, \Omega) \mathrm{d}\boldsymbol{x}. \tag{6}$$

2.6 Gain Coefficients at the FDN Input

Let Φ_{in} be the energy flux input at reverb audio input and let $\beta_n{}^2$ be the attenuation coefficient that gives the energy flux as perceived at the listener position due to 1^{st} order reflection off the n^{th} surface patch.

Recall from equation (6) that $E_1(A_n, L)$ denotes the irradiance at the listener due to 1^{st} order reflections off the patch A_n. It follows that the following relation must hold,

$$E_1(A_n, L) = (F_n \beta_n)^2. \tag{7}$$

The term ℓ_1 in equation (6) can be computed by applying the ARE in the usual way[2].

However, directly multiplying the input or output of the n^{th} channel of the FDN by β_n would yield an incorrect result because we have already multiplied υ_n at the output of each channel to model 1^{st} order reflection gain. Instead, we define μ_n to be a gain coefficient at the input of the n^{th} channel of the FDN and set it as follows,

$$\mu_n = \beta_n / \upsilon_n. \tag{8}$$

The effect of this is that for 1^{st} order reflections, the output coefficient υ_n is canceled out by the input and the resulting 1^{st} order FDN output is exactly as given in equation 7, but for second and higher order FDN output, the result is as specified by equation 5, because the signal only passes through the μ_n scaling coefficient on the first entry into the FDN; in subsequent loops it bypasses the input coefficient. See Figure 1 for a diagrammatic representation of this.

2.7 Modeling Interaural Effects

In order to model the interaural differences in timing and power spectrum that result from the orientation of the listener's ears relative to the direction of incoming acoustic rays, we use a bank of filters and delays. In Figure 1 these are labeled ITD and HRTF, which stand for Interaural Time Delay and Head-related Transfer Function.

In reality, the interaural time delay and the head related filtering effects are different for every possible angle of incidence. However, rays coming from similar angles will have similar delay times and filter transfer functions. Therefore we can approximate the interaural differences by quantising each incoming angle into M sectors around the listener's azimuth, and processing incoming acoustic rays that quantise into the same sector with the same HRTF filter and ITD delay.

To accomplish this, we place a multiplexer between the FDN and the filter and delay banks. This multiplexer mixes each of the FDN output channels into one of the M filters for the left ear and another for the right ear, according to the quantised angle of incidence from the surface geometry patch represented by the FDN channel to the listener position.

When we perform the acoustic modeling for first order reflections, we set the delay time according to the distance from each surface patch to the nearest of the listener's two ears. To compensate for the additional delay to reach the ear on the far side of the listener's head, we use the bank of inter-aural delays. The interaural delay time for a given angle is zero for the near-side ear and non-zero for the far-side ear. Since the inter-aural delay time depends only on the angle of incidence in the horizontal plane, we reduce the number of inter-aural delays by quantizing the angles of incidence into a small number of groups.

For the HRTF filterbank, we use a pole-zero filter model for a spherical head as proposed by Brown and Duda [27]. We also use the same filter for the direct rays except that for direct rays we input the exact angle of incidence without quantising. It is known that the spherical head model lacks the general boost between 2 to 7 kHz that is typically caused by ear canal and concha resonance [28] and the high frequency roll-off or notch above 8kHz depending on front-back configuration [29]. Further explanations can also be found in [30, 31]. While it is impossible to model every individual HRTF, one may add simple pole-zero EQ and

[2] We refer readers unfamiliar with this method to [26], where the authors explain it in detail.

low-pass filters at each channel or at the mixdown output of the channels, to mimic the desired general boost between 2 to 7kHz and roll-off above 8kHz. This is similar to general filtering effects applied at certain in-ear headphones that attempt to mimic the reproduction of 'live recording'. Our informal listening test indicates that the addition of these filters improve the overall quality of the simulation, while only causing negligible changes in the objective evaluation parameters.

2.8 Method Summary

In summary, the goal of the acoustic modeling calculations in this method is to set the gain coefficients at the input and output of each delay line in the network, shown in Figure 1 as μ_n and υ_n. The procedure can be outlined as follows,

1. We discretise the surface geometry $\mathcal{G}$ into a set of a total of N discrete patches $A_n \subset \mathcal{G}$, for $n = 1...N$.
2. Set the length of the n^{th} delay line to correspond to the timing of the first-order reflection that comes from the n^{th} patch of surface geometry.
3. *Compute* υ_n: Assuming that late reverb energy flux reflects diffusely and is evenly distributed over the room surface geometry, estimate the fraction of the total energy flux that should reach the listener from each of the N surface patches in the virtual room. Use the results to set the values $\upsilon_1, \upsilon_2, ...\upsilon_N$, shown in Figure 1 using equation (4).
4. Using the Acoustic Rendering Equation in [5], model the 1^{st} order reflections and compute the amount of energy at the listener due to 1^{st} order reflections using equation (6). Each of the N surface patches in our virtual room produces one first-order reflection. Each of those reflections corresponds to one delay line in the FDN.
5. *Compute* μ_n: Let β_n represent the gain of the 1^{st} order reflection. Compute it with equation (7) using values from (7) obtained in the previous step. Then the coefficient μ_n at the input of the n^{th} delay line in the FDN is $\mu_n = \beta_n/\upsilon_n$. The effect of this is that the gain of the first impulse issued from each delay d_n is exactly β_n.
6. Subsequent reflections from that same delay d_n will enter the delay line directly from the mixing matrix without passing through the input gain coefficient μ_n, hence, υ_n will at the delay output will scale the late reverb signal for that delay proportional to the energy flux output of the n^{th} patch of surface geometry that we estimated in step 1 above.

3 OBJECTIVE EVALUATION

3.1 BRIR Recordings

To evaluate the performance of our proposed method, we use BRIR samples taken from seven different rooms. Two of the impulse responses are taken from the *AIR* database [32] and we measured the others ourselves. The rooms are as follows,

- R1: A lift lobby (1.95m by 5.52m by 2.9m) in a basement. The floors and walls are made of marble, and the ceiling is made of painted concrete. There are three alcoves for lift doors which were closed during the recording. The door at the entrance is wooden. The average reverberation time of this room is 1.81s.
- R2: A long, empty, rectangular room (1.42m by 7.23m by 2.61m) with concrete walls, ceiling, and floor with three wooden doors. The room serves as an entryway for two dry riser closets. The average $RT60$ reverberation time is 1.2s.
- R3: A small, empty, almost square room (2.68m by 2.75m by 2.98m) that serves as a smoke-stop lobby to minimise the entry of smoke into the emergency staircase in the next room. There are in total of two emergency doors leading to this room, which were closed at all times. The room is made of concrete, with an average reverberation time of 2.2s.
- R4: A lecture room from the AIR database (10.8m by 10.9m by 3.5m) containing desks and chairs. The average reverberation time of this room is about 0.8s.
- R5: A meeting room from the AIR database (8m by 5m by 3.5m) with a conference table and several chairs. This room has an average reverberation time of 0.23s.
- R6: An office room from the AIR database (5.00m by 6.40m by 2.90m) with several office furnitures such as wooden desks, shelves, and chairs. The average reverberation time is 0.43s.

We measured two configurations of source and microphone positions (labeled P1 and P2) R2, and R3. Seven source-microphone configurations were measured in R1. Two representative positions (labeled P1 and P2) were selected for objective evaluation in section 3.5. The rest of the configurations were used for listening test explained in section 4.2.1 instead. For BRIRs from [32], we took two configurations in R6 and one source-microphone configuration in each of the other rooms. In total, we used 10 BRIR recordings for the objective part of the evaluation.

To measure and record BRIRs in R1 to R3, we used the logarithmic sine sweep method presented in [33]. A 50s logarithmic sweep is generated between 50Hz and 20kHz using an omni-directional speaker with sufficient volume so that the resulting BRIR has a minimum decay range of 57 dB [34]. The response of the speaker is shown in Figure 2. The signal was recorded using a pair of omni-directional binaural microphones (BE-P1) that are placed inside the ear canals of an artificial head (B1-E) which has a diameter of approximately 16.8cm. We use Lundeby's method [35] to find the point where the signal level falls below the noise floor and truncate the impulse response at that point. They are then equalized to minimize the effects introduced by the speaker response.

3.2 Implementation of the Acoustic Simulation

We implemented the proposed method in C++ in an iOS application that directly processes the input signal in real time as an algorithmic reverb. Our method can process the

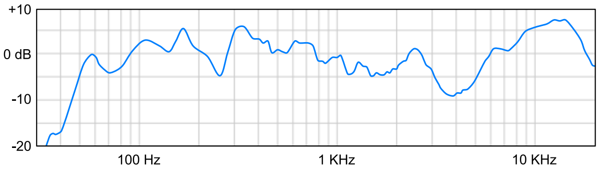

Fig. 2: The frequency response of the omni-directional speaker used to measure BRIRs in R1, R2, and R3.

input signal directly through the FDN as opposed to producing an impulse response and doing convolution because the proposed method processing directly is more efficient than a convolution reverberator. This results in much faster update times, especially in rooms with long reverberation times because it eliminates the need to produce a new impulse response of several seconds in length every time we want to update parameters. However, for the objective part of our evaluation, we did produce impulse responses using the proposed method so that we could make measurements on them.

We simulated a set of 10 BRIRs corresponding to the rooms described in the previous section using 3D meshes subdivided into 32, 64 and 128 patches. In each case, the size of the FDN corresponds to the number of patches in the mesh because the energy flux output at each patch is modeled by one channel of the FDN. With even subdivision of the mesh, this ensures that the average length of the FDN is at least as long as the mean free path of the room, as recommended in [36].

For the numerical integration, we used the Monte Carlo method with 50 sample points per mesh patch. In our interaural model, we quantised incoming angles into 12 sectors. There is some variation in the results due to the randomisation in Monte Carlo integration, so we repeated each simulation 20 times and report the average in our results section.

For comparison, we also implemented two baseline methods in C++ to simulate the two sets of 10 BRIR recordings,

1. *Baseline Method 1 (Baseline ISM)*: We generate the binaural impulse response up to third order using the ISM method [2] implemented with a multi-tap delay and send the delay tap outputs representing the third order reflections into an FDN with 32, 64, or 128 delay lines to model late reverb. This is similar to the implementation in [19].
2. *Baseline Method 2 (Baseline ARE)*: We generate the the BRIR up to second order reflections using the ARE [5] and a multi-tap delay, and route the delay tap outputs corresponding to second order reflections into an FDN with 32, 64, or 128 delay lines to model the late reverb. Corresponding to the number of delay lines in the FDN, the 3D model of the room is discretized into 32, 64, or 128 patches as well. To solve the ARE we use Monte-Carlo numerical integration with 50 points per patch. We multiplex the second order output into the FDN, such that the output from the corresponding patch is grouped together as an input to the delay line that represents reflection from that particular patch.

We used the fast Hadamard Transform to do the mixing operation for the FDN in all cases. To make a fair comparison, the FDN used in both baseline methods is identical to the FDN used in the proposed method, where the length of each delay line in the FDN is the time taken for sound to travel from the source to one of the surface patches in the room and finally to the listener. To model interaural effects in both baseline methods, we apply head-related transfer function filters and interaural time delays to each individual reflection, instead of quantising angles into sectors like the proposed method does (explained in section 2.7).

3.3 Computation Time

Since the proposed method processes directly on the input signal as an algorithmic reverb, rather than producing an impulse response for convolution, the most important measurement with respect to its performance is the time to update the model parameters following a change in listener or source position. The parameters that update with each change are the lengths of delays in the FDN, the input and output coefficients μ_n and υ_n and the multiplexer coefficients that determine to which HRTF filter and interaural delay each channel of the FDN mixes to, according to the angle between the listener and the surface geometry patch each FDN channel represents. The update times of the proposed method for three different mesh sizes are shown in Table 1. Note that when we compare the proposed method against the baseline methods, the baseline methods work by convolution rather than directly processing the input signal, so the most meaningful way to compare the two is to compare update time of the proposed method against time to render an impulse response for the baseline methods. Also note that white the update time for the baseline methods depends on the length of the impulse response but update time for the proposed method does not. The binaural early reflections units of the baseline methods are too slow for realtime processing directly on the input signal as algorithmic reverbs, so we are forced to implement them using convolution instead.

Table 1 also shows the time required to produce an impulse response for the proposed method and two baseline methods with mesh sizes of 32, 64, and 128 patches. Note that the *ISM* implementation in the baseline method assumes a rectangular room shape so it uses a fixed mesh size of 6 surfaces for early reflections but for late reverb it uses an FDN of order corresponding to the mesh size reported in the top row of the table. Rooms R1 to R6 are close to ideal rectangular shapes. We use an implementation of the *ISM* for perfectly rectangular rooms that is significantly more efficient than implementations supporting arbitrary geometry [2]. If arbitrary room shape is used, the computational time using ISM will be much longer. The ARE implementation we use is capable of supporting arbitrary room shapes and its performance depends only on the density of the mesh.

All values in Table 1 are averages of 20 simulations running on a Mac laptop with 2.5 GHz Intel Core i7 CPU and 16GB RAM on code compiled from C++. The study in [37] states that to create a realistic acoustic simulation in virtual reality systems, an update is required every 550ms when the user is navigating around the room at a normal walking speed, as the overall acoustics of a room do not drastically change for small changes in listener position. In the case of room acoustics simulations where not only direct signal but also reverberation is present, a lower update rate for the reverberation (both early and late reflections) is acceptable. Also, according to the study in [38], a latency of 80ms and below between a head-tracker and a direct audio signal is low enough so that listeners don't detect the lag. As shown in Table 1, the update time for the proposed method is less than 80ms, even for the finest mesh setting.

Update Time (ms) for Baseline and Prop. Methods

Mesh Size	32	64	128
Prop. Method Direct	11.26	24.20	49.18
B. Method ISM BRIR	176.32	243.61	351.97
B. Method ARE BRIR	8009	30830	123088
Prop. Method BRIR*	192.15	249.91	411.98

Table 1: The direct update time for the proposed method is the time it takes to re-calculate the model parameters for a change in listener or source position. The baseline methods work by convolution, hence the reported time is the time they take to render a 1.8 seconds long BRIR. For comparison, we also report the time that the proposed method would require to render a BRIR of the same length. *Please note that in implementation the proposed method never actually renders any BRIR because it is implemented as an algorithmic reverb rather than a convolution reverb.

3.4 Objective Evaluation Parameters

ISO 3381-1:2009 defines a list of parameters to measure and describe the characteristics of a BRIR, measured in the 500Hz and 1000Hz frequency bands [20]. They are reverberation time (RT_{60}), early decay time (EDT), definition (D_{50}), clarity (C_{80}), center time (T_S), and interaural correlation coefficient ($IACC_{E3}$). Except for the $IACC_{E3}$, they are all averaged between the left and right channels. We measure $IACC_{E3}$ in three octave bands: 500Hz, 1000Hz, and 2000Hz as suggested in [39] so that these values can be used to directly indicate the apparent source width.

To quantify the amount of error the simulated BRIRs has in terms of the above room parameters, we use the JND. JND is defined as the smallest amount of change in a particular variable that is noticeable more than half of the subjects of interest [40]. The JND values for RT_{60} and EDT is set as a deviation of 5% between measured and simulated values. For D_{50}, C_{80}, and T_S, it is set as 0.05, 1dB, and 0.01s absolute difference between measured and simulated values respectively. The JND values for these five room parameters are computed in the average of 500Hz and 1000Hz frequency bands. For $IACC_{E3}$, it is counted as 0.075 absolute difference between measured and simulated values in the average of 500Hz, 1000Hz, and 2000Hz frequency bands.

Since we set the RT_{60} decay time of the FDN to match the measured decay time of each room (as opposed to calculating decay time using Sabine's formula) the simulated BRIRs from the proposed and baseline methods closely match the recorded BRIR. All of the simulated IR decay times are less than 0.5 JND from the measured BRIR decay time. Therefore for the subsequent sections, we will not continue to report results for decay time.

3.5 Results

3.5.1 Comparison with Measured BRIR

Table 2 shows the raw values of all five room parameters from the measured BRIR and from the BRIR produced by the proposed method using 128 patches. Values that are greater than 1 JND are printed in bold. The evaluation in [19] does not present results from different source and listener positions in the same room. However, we feel it is relevant to take measurements at several different source and microphone positions in the same room because we observed significant position-dependent variation in some of the parameters. For example, the difference in T_s between P1 and P2 in R6 is more than 1 JND (more than 0.01s), of which both effects are captured by the proposed method using 128 patches. Among the three acoustic parameters that indicate the balance of energy between early and late reflections (D_{50}, C_{80}, and T_S), C_{80} seems to have the most cases where its error is larger than 1 JND. The absolute value of C_{80} error is also larger than both D_{50} and T_S for most of the 10 simulations. A possible reason for this is that the study in [41] recommends that the JND value for C_{80} should be 3 dB, which is three times higher than the value suggested in the ISO standard [20], which we are using to report the data in Table 2. If 3dB is used as a JND value for C_{80}, the mean JND of C_{80} for the proposed method would be below 1 JND.

The error in terms of absolute JND for EDT is 2.15 for proposed method using 128 patches, which is relatively much larger than the rest of the parameters. Table 2 also shows that the EDT values for six out of 10 simulated locations has error larger than 1 JND. In general EDT is known to be very sensitive to small errors [20]. In GA methods, we typically see wider margins of error in the EDT than other parameters. We postulate that inaccurate modeling of the bi-directional reflection function may be the cause of this. An accurate BRDF model significantly increases the computational cost of doing numerical integration. For that reason, efficient applications of the ARE typically use pure specular reflection, pure diffuse reflection, or both of them combined. None of these options is an accurate representation of the physical reality. In our implementation, the baseline *ISM* method models pure specular reflection. The proposed method and the baseline ARE method use pure diffuse reflection. In Table 2, a significantly higher EDT error is observed in R6. The proposed model actually may even actually yield higher error with a more detailed

BRIR	Measured	Prop. Method	Measured	Prop. Method	Measured	Prop. Method	Measured	Prop. Method	Measured	Prop. Method
	IACC		**D**$_{50}$		**C**$_{80}$ (dB)		**T**$_S$ (s)		**EDT** (s)	
R1 P1	0.417	**0.340**	0.317	0.295	-0.577	-1.281	0.139	0.145	1.979	2.010
R1 P2	0.335	0.393	0.345	0.320	-1.620	-0.813	0.138	0.145	2.021	2.115
R2 P1	0.226	0.203	0.469	**0.384**	1.445	**0.356**	0.094	0.095	1.594	**1.232**
R2 P2	0.305	0.345	0.455	0.422	1.173	1.777	0.102	**0.090**	1.605	**1.298**
R3 P1	0.263	0.308	0.314	0.315	-1.199	-1.607	0.150	0.154	2.059	**2.213**
R3 P2	0.246	0.305	0.317	0.322	-0.771	-1.748	0.152	0.148	2.268	2.172
R4	0.433	0.434	0.577	**0.640**	4.167	**5.170**	0.064	**0.053**	0.876	**0.950**
R5	0.722	0.665	0.947	0.969	18.483	19.439	0.016	0.009	0.166	0.166
R6 P1	0.557	0.549	0.772	0.793	9.963	**8.888**	0.031	0.029	0.559	**0.641**
R6 P2	0.778	**0.682**	0.897	0.883	12.342	11.762	0.019	0.018	0.413	**0.514**

Table 2: The values of all five room parameters of the measured BRIRs and simulated BRIRs using the proposed method with 128 patches. Results in bold are more than 1 JND from the measured result.

subdivision of the model. For example, the mean absolute error of EDT using 64 and 32 patches is 3.86 and 2.41 JND respectively. This suggests that our 3D mesh does not accurately represent the shape of that room. We obtained the impulse response for that room from the AIR database and set the parameters of the 3D model based on the description reported in [32].

3.5.2 Comparison with Baseline Methods

Prop. Method	128	64	32
IACC	0.620	**1.781**	**3.056**
D50	0.580	0.804	0.827
C80	0.820	0.675	0.930
TS	0.562	0.639	0.771
EDT	2.149	**3.856**	2.406

Table 3: Mean absolute JND values from all 10 BRIRs using proposed method, with 32, 64, and 128 of patches. Values that perform worse than either baseline methods are printed in bold.

In this section we compare the performance of the proposed method against the two baseline methods we described in section 3.2, which use a separate multi-tap delay and FDN for early reflections and late reverb. One baseline method simulates early reflections up to the 3^{rd} order using the image source method and the other uses the acoustic rendering equation up to the 2^{nd} order. We compare the performance of each using three mesh densities: 32, 64, and 128 patches. The FDN size for each method corresponds to the mesh size. Tables 5 and 6 present the mean of the absolute value of the modeling error for the baseline ARE and ISM methods, respectively, in units of JND. In Table 3, the mean absolute error value of the proposed method is printed in bold when it is greater than either one of the baseline methods and in plain text when it is less than both of them. Note that for mesh sizes 64 and 32, the proposed method performed worse on IACC than the baseline methods. Recall that in the implementation of the interaural effects of the baseline methods, the HRTF and ITD filters are applied to each individual reflection, while in the proposed method we have only an eight channel filterbank. This may imply that the error introduced by quantising the angle can be compensated with a finer mesh setting. The errors for EDT in all three methods are significantly higher than the rest of the room parameters. The authors in [42] report that EDT is sensitive to changes in scattering coefficients. The FDN used in the proposed method mixes energy in equal amounts from each patch in the room to every other patch. This does not correspond to any physically informed model of scattering. Based on the work presented in [19] and [43] it appears that in general, hybrid geometrical acoustic simulation methods do not model EDT well.

We also conducted a Wilcoxon Signed-Rank test to compare the performance of the proposed and baseline methods. We compare the mean absolute error of the proposed method against each of the two baseline methods. Since the proposed method uses the ARE to model only the 1^{st} order reflections, we do not intend for it to out-perform either of the baseline methods, which model early reflections up to second order using the *ARE*, or third order using the *ISM*. Our goal is only to have the proposed method achieve close to accuracy of the baseline while being significantly more efficient. See Table 1 for timing data.

In the Wilcoxon test, our alternative hypothesis H_α is $(|\mu_{\text{prop}}| - |\mu_{\text{baseline}}|) < 1$, where $|\mu|$ represents the mean absolute JND of all 10 simulated BRIRs. In other words, the alternative hypothesis states that the difference between the absolute value of mean of the proposed method and the baseline method is less than 1 JND. The motivation for this hypothesis is that we want to show that the proposed method, although simpler and faster than the baseline methods, is not audibly less accurate.

Table 4 shows the p-values of the test. Except for IACC, all of the simulation results support rejecting the null hypothesis with at least 99% confidence level. This shows that despite being simpler and more efficient than the baseline method, the average simulation error of the proposed

	Mesh Size 32		Mesh Size 64		Mesh Size 128	
Baseline Method	ISM	ARE	ISM	ARE	ISM	ARE
$IACC_{E3}$	0.500	0.500	0.216	0.053	0.001*	0.002*
D50	0.001*	0.007*	0.002*	0.002*	0.003*	0.002*
C80	0.005*	0.014*	0.002*	0.001*	0.001*	0.001*
TS	0.003*	0.003*	0.001*	0.001*	0.000*	0.013*
EDT	0.010*	0.014*	0.216	0.001*	0.024*	0.001*

Table 4: The p values for Wilcoxon Signed-Rank test with H_α: $|\mu_{\text{prop}}| - |\mu_{\text{baseline}}| < 1$, where $|\mu|$ represents the mean absolute JND, testing against different baseline methods: the ISM and ARE baseline methods for mesh sizes 32, 64, and 128. p-vals in asterisk (*) are those that are less than 0.05, indicating tests that have confirmed the alternate hypothesis at 95% confidence level.

B. Method ARE	128	64	32
IACC	3.116	2.284	1.380
D50	1.029	1.018	0.827
C80	1.670	1.334	1.489
TS	1.089	0.977	1.193
EDT	3.566	2.781	2.641

Table 5: Mean absolute JND values from all 10 BRIRs using baseline ARE method (named as B. Method ARE in the table), with 32, 64, and 128 of patches.

B. Method ISM	128	64	32
IACC	3.065	1.543	1.560
D50	0.865	1.051	1.001
C80	1.753	1.051	1.001
TS	1.061	0.927	1.094
EDT	4.446	4.009	4.644

Table 6: Mean absolute JND values from all 10 BRIRs using baseline ISM method (named as B. Method ISM in the table), with 32, 64, and 128 of patches.

method is less than 1 JND higher than the baseline methods. For IACC we can reject the null hypothesis only for the size 128 mesh. This supports our conjecture that modeling a perceptually accurate IACC requires some minimum amount of acoustic rays per square meter. The result might also imply that simulation of higher order reflections improves the accuracy of IACC when the number of patches used is small.

It is worth noting that it is possible to model second order reflections using the proposed method. That could be implemented by simulating second order reflections using the ARE and using the results to set the FDN delay times and output gains in exactly the same way that we do with the first order reflections. To test that idea, we implemented that method of second order modeling in the proposed method and ran some informal tests. We found that it increased update time with very little improvement in the accuracy. Since we intend for the proposed method to be efficient rather than accurate, we do not include those results in this paper.

4 SUBJECTIVE EVALUATION

Our intended applications for the proposed method are virtual reality and gaming, fields where perceptual plausibility may be as important than the objective measures discussed in the previous section. In this section we evaluate our method in terms of the following five perceptual qualities: naturalness, reverberation, coloration, metallic character, and source width as suggested in [44]. The procedure and result is presented in sections 4.1.1 and 4.1.2, respectively. Additionally, we conducted a second listening test to measure the sense of spatial location that listeners perceive when listening to sounds processed through the proposed reverberator. The procedure and result for the second listening test is presented in section 4.2.1 and 4.2.2 respectively.

4.1 Part I: Listening Test Evaluation of Standard Perceptual Qualities

4.1.1 Test Subjects and Procedure

19 subjects (12 female, 7 male) with ages ranging from 20 to 40 participated in this listening test. 15 out of 19 subjects are experienced musicians. All of them reported normal hearing ability. The listening test was conducted in a small, carpeted, and enclosed meeting room. The room was quiet as its air conditioner was switched off to further eliminate background noise. The test was delivered using a pair of AKG-702 headphones and a headphone amplifier at a sampling rate of 44.1 kHz.

For the listening test, we selected four representative BRIRs (R2 P1, R3 P1, R4, and R6) from the 10 BRIRs we used in the section 3. They are selected such that we have a variation in both room size and reverberation time. Two 8s long anechoic input signals, a male spoken speech and a guitar piece were convolved with both the measured BRIR and the synthesized BRIR using 64 and 128 patches. Also, since geometric acoustics methods do not accurately simulate wave phenomena in the lowest frequencies of the audio band [45], we filtered the dry audio signals to exclude frequencies below 100 Hz. To ensure fair comparison across

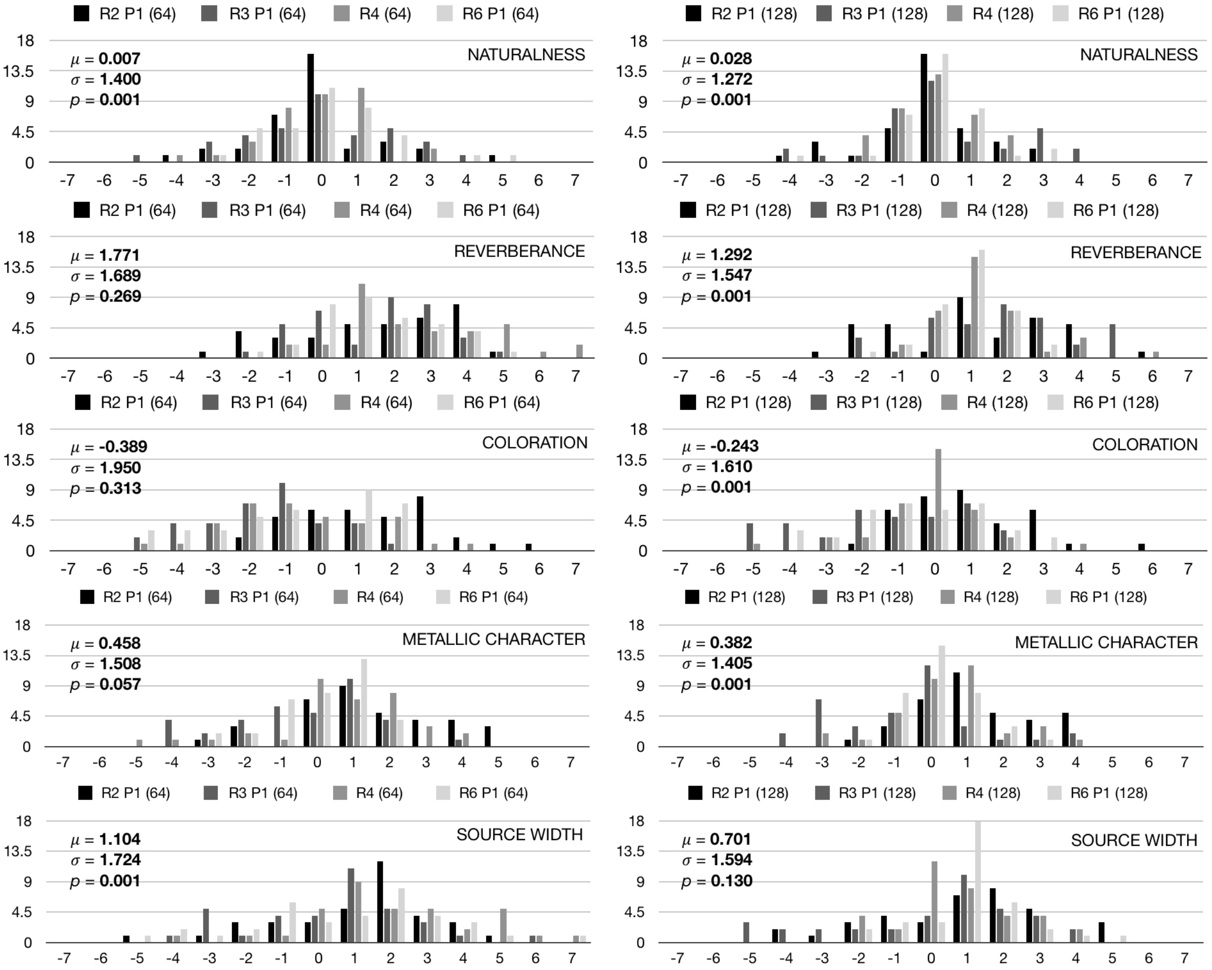

Fig. 3: Histogram of the 15-scale bipolar ratings by 19 subjects on all five perceptual qualities, using the synthesized signal from the proposed method with 64 patches (left) and 128 patches (right). The rating scale is explained in section 4.1.1. The mean (μ) and standard deviation (σ) of the rating across all rooms and subject is presented for each histogram. In each sub-figure we also show the p value obtained from the Lilliefors test.

listening test subjects of various ages, we also filtered out frequencies above 15 kHz as recommended in [46].

We presented each listener with sets of three audio files at a time, one file convolved with the measured BRIR and two more processed with the proposed method using 64 and 128 patches in the mesh. We refer to these three types of samples as *measured signal*, *synthesized signal 64* and *synthesized signal 128*.

Each subject was asked to compare the degree of naturalness (less - more), reverberance (less - more), coloration (darker - brighter), metallic character (less - more), and source width (smaller - larger) of the synthesized signals to the measured signals and rate each of them on a 15-point bipolar scale (anchored at -7 and 7 for both extreme ends). This is a general scale often used for subjective tests of perceptual qualities. It has been shown to produce reliable results and reduce grade inflation [47, 48].

The descriptions for the ratings given to the subjects are as follows: 0 for exactly the same, 1 or -1 for similar, 2 or -2 for very slightly different, 3 or -3 for slightly different, 4 or -4 for moderately different, 5 or -5 for quite different, 6 or -6 for significantly different, and 7 or -7 for extremely different. The order of the five perceptual qualities to be rated by each subject was randomized.

To prevent exhaustion, we encouraged the subjects to take small breaks in between and take as much time as they want in completing the test. The subjects took between 45 and 60 minutes to comfortably finish the test.

4.1.2 Results

Figure 3 shows histograms of the ratings given by all 19 subjects, in all four locations for the *synthesized signal 64* (left) and *synthesized signal 128* (right), with two samples rated at each location. Each histogram represents a total of 152 ratings. The mean and standard deviation on the ratings across all rooms by all 19 subjects are shown beside each histogram. We also show the p value obtained from Lilliefors test to indicate the normality of the dataset. Given the limited number of participants, we do not always

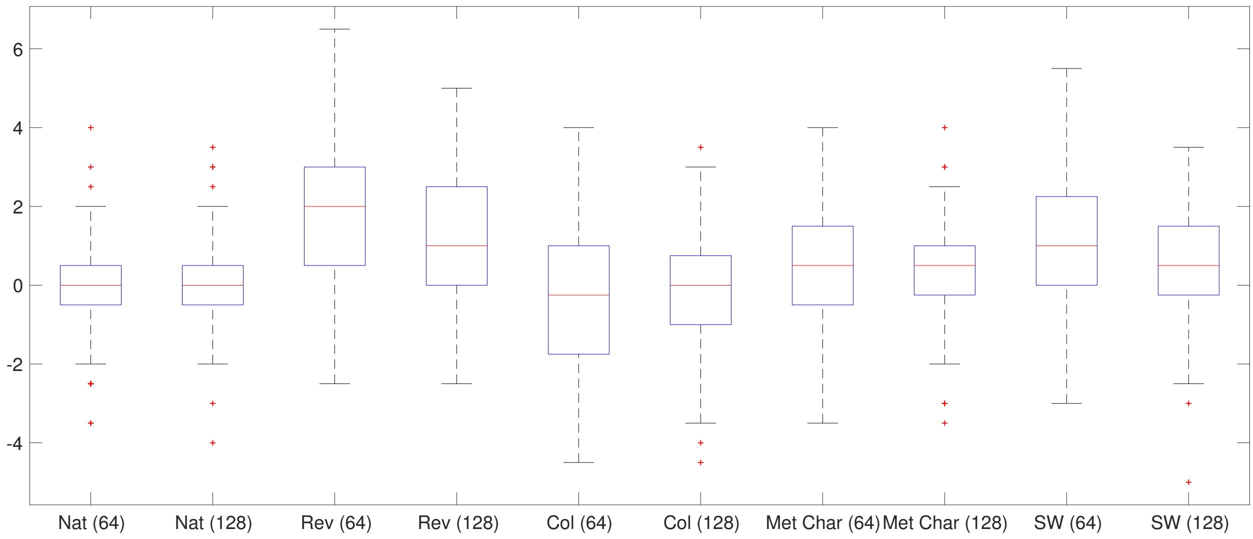

Fig. 4: Boxplots of the 15-scale bipolar ratings by 19 subjects on all five perceptual qualities: Naturalness (Nat), Reverberance (Rev), Coloration (Col), Metallic Character (Met Char), and Source Width (SW) using 64 and 128 patches. Each boxplot contains 76 responses in total from 19 subjects and 4 different room condition.

expect normality in the response. However in general, the results show a fair consistency between measured and synthesized signals, as each histogram has single peak with roughly equal amount of variance on each side. Figure 4 shows the boxplots of the same dataset. Most answers are roughly symmetric about the median, and the median of the dataset is close to the mean for each condition.

As expected, synthesized signals using 128 patches are rated as perceptually closer to the measured signals as compared to synthesized signals using 64 patches. We conducted a Wilcoxon Signed-Rank test to validate this claim. The alternate hypothesis is that the absolute rating using 128 patches is lesser than the absolute rating using 64 rating. With 5% significance level, we found that the p-values are 0.031, 0.001, 0.001, 0.032, and 0.102 for naturalness, reverberance, coloration, metallic character, and source width respectively.

Most subjects rated the synthesized signal as exactly as natural as the measured signal. In general, subjects viewed the synthesized signal as more reverberant than the measured signal. This contradicts the fact that the reverberation time between measured and synthesized signal is always less than 0.5 JND, suggesting that they shouldn't be noticeable at all. We noticed that the main weakness of synthesized signal with 64 patches appears to be coloration, with most subjects gave negative rating, and it also has a larger spread as compared to the rest of the histograms. The response looks like a bimodal distribution. There is also a slight error in the perception of source width, where most subjects rated both synthesized source widths as larger than the measured ones. This error might be attributed to the fact that the synthesized method only use simple spherical-head approximation as HRTF.

4.2 Part II: Measuring the Sense of Spatial Location

4.2.1 Test Subjects and Procedure

The goal of the the second part of our subjective evaluation is to determine how effectively the proposed method generates perceptual cues that allow listeners to determine their position in a virtual room. To do this, we conducted listening tests where we showed listeners several images with the listener and sound source locations marked on the map of a room and asked them to select the image that best corresponded to their auditory perception. The tests described in this section attempt to answer the question, *does the loss of detail resulting from a rough and simplified approximation (like the method proposed here) negatively affect the listener's ability to perceive his or her own location and the spatial characteristics of the room?*

We conducted tests with 11 experienced listeners (4 females and 7 males), all reported normal hearing ability. The test subjects include one recording engineer, five virtual-reality gamers who report familiarity with listening to spatial audio localisation cues, and five academic researchers in audio-related fields. 6 out of 11 subjects are musicians. The age of test subjects ranges from 26 and 40 years. Each test took between 25 to 40 minutes to complete, and we conducted them using the same hardware: a MacBook pro, a vacuum tube headphone amplifier, and a set of AKG Q-701 headphones. The test was carried out in a quiet environment as the one described in part I of the listening test, therefore there was negligible background noise and it imposed no effect on the results.

To produce the recordings used in the listening test, we obtained a recording of an acoustic guitar recorded with the microphone up close with no audible room reverberation [49]. For each configuration of listener and source position in the test, we produced two versions of the recording, one convolved with a simulated impulse response and the other with a measured impulse response. The impulse responses are taken from R1 with various source-microphone configuration.

Each question of the test consists of a pair of sound recordings and a pair of pictures showing the floor plan of a room with listener and sound source locations marked. Figure 5 shows a sample of two such questions. The complete test consists of ten questions of this type. Each listening test candidate answered the same set of ten questions twice, once with the reverb using the measured impulse response and once with the simulated reverb. We randomised the order so of the tests to eliminate the possibility that the measured IR test affected the results of the simulated IR test or vice versa. Test subjects were allowed to replay the recordings as many times as they needed. We counted the number of correct answers in each of the two sets of 10 questions from each participant.

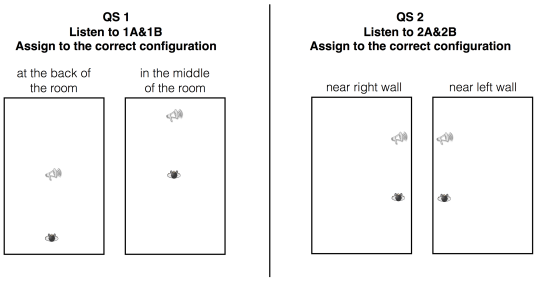
QS 1
Listen to 1A&1B
Assign to the correct configuration
at the back of the room
in the middle of the room
QS 2
Listen to 2A&2B
Assign to the correct configuration
near right wall
near left wall

Fig. 5: Sample listening test question

In the design of the listening test, we were careful to avoid posing questions where the the listener would be able to guess the correct answer on the basis of the angle between source and listener alone. For example, if we present the listener with a question where answer choice A showed a source-listener configuration where the source is to the left of the listener and choice B showed the source to the right of the listener, the listener could easily match the sounds to the correct room map image based on the relative volume between the left and right ears alone, without listening to the reverb at all. To ensure that we were testing the listener's perception of the reverberation rather than the direct sound from source to listener, we kept the listener and source at the same distance and angle relative to each other; the two moved around the room as a pair. Figure 5 illustrates an example of this. Therefore, any detected change in direct-to-reverberant ratio is purely due to the reverberant part of the impulse responses.

Since we used BRIRs from the same room R1, we eliminated the possibility that the listener could guess the answer based on reverberation time or other properties inherent to the room but not unique to the listener's position in the room.

4.2.2 Result

Figure 6 summarizes the normalized score of the listening test from the 11 test participants. The test has ten questions for the proposed method and ten questions for the measured impulse responses. We normalised scores onto the range $[0, 1]$, so that 1 indicates 10 out of 10 questions correct. The average score for the measured IR is (0.72) and for the simulated IR is (0.764).

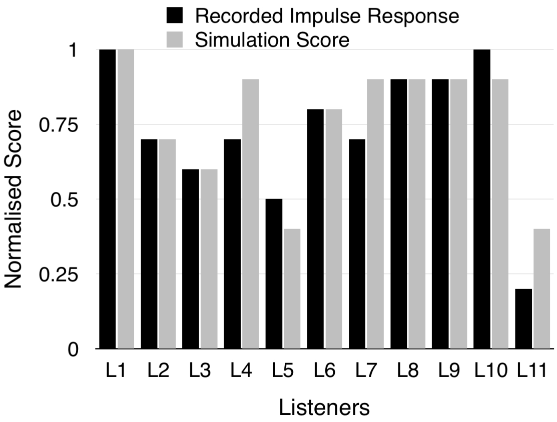

Fig. 6: Normalised listening test scores of 11 test participants, comparing results for measured (black) and simulated (grey) reverb impulse responses.

In general, candidates that scored well on the first set of questions also scored well on the second set of 10 questions, regardless of whether they used the recorded or simulated reverb first. To quantify this, we calculated the Spearman Correlation Coefficient of the two sets of results. The R-value for correlation between measured and simulated test results is 0.853, and the two-tailed value of p is 0.00085, indicating a statistically significant positive linear correlation between the score on the simulated reverb test and the score on the measured reverb test.

To investigate whether the correct answer rate is significant, we conducted a one-sided binomial test. The total sample size from all 11 listeners is 110, as each listener has to listen to perceptual cues in 10 different configurations. According to [50], for 5% significance level, the amount of correct answer percentage should be higher than 58.32% such that it is safe to assume that the answers given by the listeners were due to audible differences and not due to chance. The correct answer percentage using our method is 76.3%. This indicates that the correct answer rate is significant and that the proposed method effectively generate perceptual cues that allow listeners to determine their position in the virtual room.

The more important insight to be gleaned from the data is that with regards to their sense of auditory-spatial location, human listeners are sensitive only to the grossest and most obvious auditory cues. This is significant because it implies that our efforts to make very accurate acoustic models may be in vain if the end goal is simply to give the listener a plausible sense of spatial location. We strongly recommend further research to determine the relative perceptual importance of each of the types of auditory cues typically simulated in reverbs of this type. The method proposed here, although simpler than previous methods, was designed to maintain a reasonable level of accuracy in terms of the objective measures discussed in the previous section. If it should turn out that this level of realism is perceptually irrelevant, we might further simplify the design.

4.3 Discussion

We noted the following observations when conducting both the listening tests.

First, most of the subjects who participated in the second listening test experienced fatigue after competing both sets of 10 questions. In total, they had to listen to 40 versions of the same classical guitar recording played through convolution with 40 different reverb impulse responses (2 files per question, two sets of 10 questions). In most cases listeners chose to listen to the audio samples for each question several times. Listener fatigue may have reduced the accuracy of the test results in part II.

Second, in informal preliminary tests we tried several different headphones and found that the spatial cues became significantly clearer when using professional-standard headphones. We had difficulty discerning location when listening with the white ear-bud headphones that come included with one of the most popular brands of mobile phone. It may be worth investigating this further before deploying binaural reverb in virtual reality gaming applications because the majority of users would likely be using inexpensive headphones. Results were much better with over-ear style headphones such as the AKG-701 that selected for the listening tests.

Third, we noticed that test candidates were easily confused if they listened to a single sound file for too long. Best results were obtained when the candidates rapidly switched between the measured and synthesized signals for part I of the listening test and between A and B sound files (see Figure 5) for part II of the listening test to listen for differences, rather than listening to an entire file before switching to the other alternative.

Especially interesting feedback from the perspective of producing minimally simple perceptually plausible simulation is the listening test results, wherein many expert listeners found it easier to guess their position in a virtual room from the sound of the simulated reverb than when listening to audio processed through the real room impulse responses. First, this suggests that the human ability to perceive details in acoustic models is somewhat limited, and therefore there is no need to develop more complicated and accurate models unless the goal of the modeling extends beyond perceptual plausibility. Second, it may be that the simplified geometric models used in our listening tests resulted in clearer perceptual cues than the more complex geometry of the real spaces due to lack of distracting details. The idea that simplifying the model could actually clarify the perceptual impression is an interesting possibility that could lead to even more efficient implementations. Towards that end, it would be helpful to investigate the proposed method and other related methods piece by piece, using listening tests to determine the perceptual importance of the various pieces of the design.

5 CONCLUSION AND FUTURE WORK

The key advantages of the proposed method are simplicity and efficiency. The proposed method can directly process input signal as algorithmic reverb, and this significantly reduces its computational time because it does not need to produce an impulse response after every parameter update. The proposed method is slightly less accurate than the baseline methods we compared it with, which represent typical existing efficient binaural simulation methods. However, we showed that the difference in accuracy between the proposed and baseline methods in terms objective room parameters is mostly less than 1 JND, so by definition, the difference is not perceptible. The update time associated with our proposed method is an order of magnitude faster and it is less complex to implement on account of having a smaller number of components. In listening tests, we found a good average agreement between measured and simulated signals in terms of five perceptual qualities: naturalness, reverberance, coloration, metallic character, and source width. We also found no significant difference between the proposed method and using measured binaural impulse responses, in terms of listener's ability to guess their location in a room based on auditory cues alone. Therefore this method may be an excellent choice for applications where a more efficient method of generating perceptually plausible binaural reverb is needed.

Over the course of this project we identified several areas for future investigation related to this subject. First, the average score of listeners trying to guess their location in a room based on auditory cues alone was slightly higher with the proposed method than with measured impulse responses. It would be truly surprising if listeners actually localised better when listening to a rough approximation like the proposed method than when listening to reverb generated from real measured impulse responses. Hence, it might be helpful to do further investigation into which aspects of the simulation contribute most significantly to listener's ability to perceive their own position in a virtual room. In particular, it would be especially useful to know if the listener's perception of location actually becomes clearer when insignificant details are removed from the impulse response.

Another interesting area for further investigation is the headphone quality issue. When deploying similar methods in user applications such as mobile gaming, the majority of users will be listening on the ear buds that come bundled with their mobile-phone purchase. Two questions arise related to this issue. First, to what extent can listeners hear localisation cues with those low cost headphones? If differences cannot be perceived then perhaps we should further simplify the binaural model to avoid wasting computational power on inaudible details. The second relevant question is, can the spatial-auditory cues be exaggerated in some way to make them easier to be perceived?

We also discovered a discrepancy in the way the proposed method models the balance of energy between early reflections and late reverb. This is significant because it affects not only the proposed method but also both of the baseline methods presented in this paper and also most of the existing methods that generate late reverb using either an FDN or convolution with an impulse response that is not specifically modeled for the particular room we are simulating. In [23], we present a detailed explanation of the error and proposed methods for correcting it. After applying the corrections proposed in that publication we repeated the series of tests shown in section 3 and observed significant improvements. We expect that similar improvements are possible with many other related methods.

6 REFERENCES

[1] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel. Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1421–1448, July 2012. https://doi.org/10.1109/TASL.2012.2189567.

[2] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small room acoustics. *The Journal of the Acoustical Society of America*, 64(943):943–950, 1979. https://doi.org/ 10.1121/1.382599.

[3] L. Savioja and U. P. Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(708), 2015. https://doi.org/10.1121/1.4926438.

[4] S. Pelzer, L. Aspock, D. Schroder, and M. Vorlander. Interactive Real-Time simulation and auralization for modifiable rooms. *Building Acoustics*, 21(1):65–73, March 2014. http://dx.doi.org/10.1260/1351-010X.21.1.65.

[5] S. Siltanen, T. Lokki, S. Kiminki, and L. Savioja. The room acoustic rendering equation. *The Journal of the Acoustical Society of America*, 122(3):1624–1635, September 2007. https://doi.org/10.1121/1.2766781.

[6] J. M. Jot and A. Chaigne. Digital delay networks for designing artificial reverberators. In *90th Convention of the Audio Engineering Society*, number 3030, February 1991.

[7] R. O. Duda, V. R. Algazi, and D. M. Thompson. The use of Head-and-Torso models for improved spatial sound synthesis. In *113 Audio Engineering Society Convention*, number 5712, October 2002.

[8] Vesa Välimäki, Julian Parker, Lauri Savioja, Julius O. Smith, and Jonathan Abel. More than 50 years of artificial reverberation. In *60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, number K-1, January 2016.

[9] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen. Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47(9):675–705, September 1999.

[10] L. Savioja, T. Lokki, and J. Huopaniemi. Auralization applying the parametric room acoustic modeling technique - the DIVA auralization system. In *Proc. International Conference on Auditory Display 2002*, July 2002.

[11] D. Schröder. *Physically based real-time auralization of interactive virtual environments*. PhD thesis, Fakultät für Elektrotechnik und Informationstechnik der Rheinisch-Westfälischen Technischen Hochschule Aachen, February 2011.

[12] D. Schroder and M. Vorlander. Hybrid method for room acoustic simulation in real-time. In *Proceedings on the 20th International Congress on Acoustics (ICA)*, 2007.

[13] M. A. Gerzon. Periphony: With-Height sound reproduction. *Journal of the Audio Engineering Society*, 21(1):1–2, February 1973.

[14] D. Schröder, P. Dross, and M. Vorländer. A fast reverberation estimator for virtual environments. In *30th International Conference: Intelligent Audio Environments*, number 13, March 2007.

[15] D. Schröder and T. Lentz. Real-Time processing of image sources using binary space partitioning. *Journal of the Audio Engineering Society*, 54(7/8):604–619, July 2006.

[16] F. Menzer. Efficient binaural audio rendering using independent early and diffuse paths. In *132 Audio Engineering Society Convention*, number 8584, April 2012.

[17] F. Menzer and C. Faller. Binaural reverberation using a modified jot reverberator with Frequency-Dependent interaural coherence matching. In *126 Audio Engineering Society*, number 7765, May 2009.

[18] F. Menzer. Binaural reverberation using two parallel feedback delay networks. In *Audio Engineering Society 40th International Conference: Spatial Audio: Sense the Sound of Space*, October 2010.

[19] T. Wendt, S. Van de Par, and S. Ewert. A Computationally-Efficient and Perceptually-Plausible algorithm for binaural room impulse response simulation. *Journal of the Audio Engineering Society*, 62(11):748–766, December 2014. https://doi.org/10.17743/jaes.2014.0042.

[20] ISO 3382-1:2009 - acoustics – measurement of room acoustic parameters – part 1: Performance spaces. Technical report, 2009.

[21] Hequn Bai, Gael Richard, and Laurent Daudet. Geometric-based reverberator using acoustic rendering networks. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*, pages 1–5. IEEE, October 2015. http://dx.doi.org/10.1109/WASPAA.2015.7336934.

[22] D. Griesinger. Spaciousness and envelopment in musical acoustics. In *Proceedings of 101st Audio Engineering Society Convention*, number 4401, 1996.

[23] H; Anderson, N; Agus, J. M; Chen, and S. Lui. Modeling the proportion of early and late energy in Two-Stage reverberators. *Journal of the Audio Engineering Society*, 65(12):1071–1031, December 2017. https://doi.org/ 10.17743/jaes.2017.0041.

[24] J. M. Jot. Efficient models for reverberation and distance rendering in computer music and virtual audio reality. In *Proc. 1997 International Computer Music Conference*, 1997.

[25] H. Anderson, K. W. E. Lin, C. So, and S. Lui. Flatter frequency response from feedback delay network re-

verbs. In *41st International Computer Music Conference 2015*, September 2015.

[26] N. Agus, H. Anderson, J. M. Chen, and S. Lui. Energy-Based binaural acoustic modeling. Technical Report 1, Singapore University of Technology and Design, April 2017. https://istd.sutd.edu.sg /research /technical-reports/energy-based-binaural-acoustic-modeling.

[27] C. P. Brown and R. O. Duda. A structural model for binaural sound synthesis. *IEEE Transactions on Speech and Audio Processing*, 6(5):476–488, September 1998. http://dx.doi.org/10.1109/89.709673.

[28] E. A. Shaw and R. Teranishi. Sound pressure generated in an external-ear replica and real human ears by a nearby point source. *The Journal of the Acoustical Society of America*, 44(1):240–249, July 1968. https://doi.org/10.1121/1.1911059.

[29] S. Carlile. *Virtual Auditory Space: Generation and Applications*. Springer Science & Business Media, 2013.

[30] B. B. Ballachanda. Theoretical and applied external ear acoustics. *Journal of the American Academy of Audiology*, 8(6):411–420, December 1997.

[31] P. A. Hellstrom and A. Axelsson. Miniature microphone probe tube measurements in the external auditory canal. *The Journal of the Acoustical Society of America*, 93(2):907–919, February 1993.

[32] M. Jeub, M. Schäfe, and P. Var. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proceedings of the 16th International Conference on Digital Signal Processing*, DSP'09, pages 550–554, Piscataway, NJ, USA, July 2009. IEEE Press. https://doi.org/ 10.1109/ICDSP. 2009.5201259.

[33] A. Farina. Simultaneous measurement of impulse response and distortion with a Swept-Sine technique. In *108th Convention of the Audio Engineering Society*, number 5063, February 2000.

[34] C. C. J. M. Hak, R. H. C. Wenmaekers, and L. C. J. van Luxemburg. Measuring room impulse responses: Impact of the decay range on derived room acoustic parameters. *Acta Acustica united with Acustica*, 98(6):907–915, January 2012. https://doi.org/10.3813/AAA.918574.

[35] A. Lundeby, T. E. Vigran, H. Bietz, and M. Vorländer. Uncertainties of measurements in room acoustics. *Acta Acustica united with Acustica*, pages 344–355, July 1995.

[36] J. O. Smith. *Physical audio signal processing for virtual musical instruments and audio*. W3K Publishing, 2010.

[37] T. Lentz. *Binaural technology for virtual reality*. PhD thesis, RWTH Aachen University, Aachen, Germany, 2007.

[38] D. S. Brungart, B. D. Simpson, and A. J. Kordik. The detectability of headtracker latency in virtual audio displays. In *International Conference on Auditory Display*, volume 73, 2005.

[39] T. Hidaka, L. L. Beranek, and T. Okano. Interaural crosscorrelation, lateral fraction, and low and high-frequency sound levels as measures of acoustical quality in concert halls. *The Journal of the Acoustical Society of America*, 98(2):988–1007, August 1995. http://dx.doi.org /10.1121/ 1.412847.

[40] G. Fechner. *Elements of psychophysics*, volume 1. Holt Rinehart Winston, New York, 1966. Translation of ”Elemente der Psychophysik”.

[41] M. C. Vigeant, R. D. Celmer, C. M. Jasinski, M. J. Ahearn, M. J. Schaeffler, C. B. Giacomoni, A. P. Wells, and Caitlin I. Ormsbee. The effects of different test methods on the just noticeable difference of clarity index for musica). *The Journal of the Acoustical Society of America*, 138(1):476–491, July 2015. http://dx.doi.org/10.1121/1.4922955.

[42] L. M. Wang, J. Rathsam, and S. Ryherd. Interactions of model detail level and scattering coefficients in room acoustic computer simulation. In *Proceedings of the International Symposium on Room Acoustics: Design and Science, RADS 2004*, April 2004.

[43] R. A. Tenenbaum, T. S. Camilo, J. C. B. Torres, and L. T. Stutz. Hybrid method for NumericalSimulation of room Acoustics:Part 2 validation of theComputational code RAIOS 3. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 29(2):222–231, 2007. http://dx.doi.org/10.1590/S1678-58782007000200013.

[44] A. Lindau. Spatial audio quality inventory (SAQI). test manual. v1.2. Technical report, TU Berlin, March 2015.

[45] S. Siltanen, T. Lokki, and L. Savioja. Rays or waves? understanding the strengths and weaknesses of computational room acoustics modeling techniques. In *Proc. Int. Symposium on Room Acoustics (ISRA)*, August 2010.

[46] P. G. Stelmachowicz, K. A. Beauchaine, A. Kalberer, and W. Jesteadt. Normative thresholds in the 8- to 20-kHz range as a function of age. *Journal of the Acoustical Society of America*, 86(4):1384–1391, 1989.

[47] S. Chaiken and A. H. Eagly. Communication modality as a determinant of persuasion: The role of communicator salience. *Journal of Personality and Social Psychology*, 45(2):241–256, 1983.

[48] S. C. South, T. F. Oltmanns, and E. Turkheimer. Interpersonal perception and pathological personality features: Consistency across peer groups. *Journal of Personality*, 73(3):675–692, June 2005. https://doi.org/10.1111/j.1467-6494.2005.00325.x.

[49] M. Woirgard, P. Stade, J. Amankwor, B. Bernschütz, and J. Arend. Cologne university of applied sciences - anechoic recordings, 2012.

[50] L. E. Harris and K. R. Holland. Using statistics to analyse listening test data: some sources and advice for non-statisticians. In *Proceedings of the 25th Conference on Reproduced Sound: The Audio Explosion, Institute of Acoustics*, volume 31, pages 294–309, 2009.

THE AUTHORS

Natalie Agus

Hans Anderson

Jer-Ming Chen

Simon Lui

Dorien Herremans

Natalie Agus was born in Jakarta, Indonesia, in 1991. She is currently a Ph.D. student at the Singapore University of Technology and Design in the area of audio engineering. Her primary research interests are room acoustics and binaural room simulation.

●

Dr. Hans Anderson is director of Blue Mangoo, a software company that makes musical apps for iOS. He is currently enrolled as a Ph.D. student at Singapore University of Technology and Design. His professional interests include project management and UI design.

●

Dr. Jer-Ming Chen is assistant professor at the Singapore University of Technology and Design (SUTD), following his previous appointment as Australian Post-Doctoral (APD) research fellow (Australian Research Council) at the Music Acoustics Laboratory, University of New South Wales in Sydney, Australia. His primary research interest is music acoustics and speech science. A believer in science communication, Jer-Ming has also written lay-language scientific papers and has featured in the international media and popular press, including newspapers (e.g. New York Times, UK Telegraph, Sydney Morning Herald), TV, and radio documentaries (BBC, ABC, Network Ten), and popular science magazines (Physics Today, Scientific American, The Straight Dope).

●

Dr. Simon Lui is an Assistant Professor at Singapore University of Technology and Design (SUTD). He received his Ph.D. degree in Computer Science from The Hong Kong University of Science and Technology (HKUST) in 2011. His primary research interests are audio engineering and semantic audio information retrieval. He has 7 inventions on the iOS platform, including several 1 best selling apps in Hong Kong, Taiwan, Malaysia, Indonesia and Canada iOS app store.

●

Dr. Herremans is an Assistant Professor at Singapore University of Technology and Design (SUTD), has a joint-appointment at the Institute of High Performance Computing, A*STAR, and is Director of SUTD Game Lab. Before joining SUTD, she was a Marie Sklodowska-Curie Postdoctoral Fellow at the Centre for Digital Music at Queen Mary University of London. She received her Ph.D. in Applied Economics at the University of Antwerp. Dr. Herremans' research interests include the intersection of machine learning, optimisation and novel application such as digital music.