SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Real-Time Binaural Auralization

Submitted by

Natalie AGUS

Thesis Advisor

Dr. Simon LUI
Dr. Dorien HERREMANS

INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

A thesis submitted to the Singapore University of Technology and Design in fulfillment of the requirement for the degree of Doctor of Philosophy

2018

PhD Thesis Examination Committee

TEC Chair:	Prof. Zhou Jianying
Main Advisor:	Dr. Dorien Herremans
Co-advisor(s):	Dr. Simon Lui
Internal TEC member 1:	Dr. Jer-Ming Chen
Internal TEC member 2:	Dr. Hyowon Lee

Declaration

I hereby confirm the following:

- I hereby confirm that the thesis work is original and has not been submitted to any other University or Institution for higher degree purposes.
- I hereby grant SUTD the permission to reproduce and distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created in accordance with Policy on Intellectual Property, clause 4.2.2.
- I have fulfilled all requirements as prescribed by the University and provided 1 copy of my thesis in PDF.
- I have attached all publications and award list related to the thesis (e.g. journal, conference report and patent).
- The thesis does / does not (delete accordingly) contain patentable or confidential information.
- I certify that the thesis has been checked for plagiarism via turnitin/ithenticate. The score is 100%.

Name and signature: Natalie Agus

Date: 01 / 07 / 2018

Abstract

INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

Doctor of Philosophy

Real-Time Binaural Auralization

by Natalie AGUS

Auralization is a process to render acoustic phenomena in a given virtual space. In the literature, it is often called as room acoustic rendering or room simulation. The result of the simulation can be categorized into either monaural or binaural auralization. In monaural auralization, we simulate how acoustic waves emitted from one or more sources behave at a particular location in the virtual space. Binaural reproduction of room acoustics is a tougher task, because it also simulates how these sources actually sounds like to a human listener. In other words, it gives a three-dimensional listening experience. More processing steps that have to be done include the simulation of the effect of head, shoulder, and pinnae shadowing. Room simulation algorithms are mainly utilized in virtual reality systems that are used for gaming, music production, entertainment, training, conditioning, or study of an acoustic space. The required level of accuracy heavily depends on the context of application. We may trade-off some level of accuracy in order to achieve faster computational time. The idea of designing an algorithm that is both accurate and fast remains open for research. In this study we propose a perceptually convincing and extremely efficient model that can reproduce binaural room acoustics even on mobile devices in real-time.

Publications

Technical Report:

Agus, N., Anderson, H., Chen, J.M., Lui, S., "Energy-based binaural acoustic modeling," Singapore University of Technology and Design, Tech Report No. 1, (Apr 2017). https://istd.sutd.edu.sg/ research/technical-reports/energy-based-binaural-acoustic-modeling.

Peer-Reviewed Journals:

Anderson, H.*, Agus, N.*, Chen, J.M, Lui, S., "Modeling the proportion of early and late energy in two-stage reverberators," Journal of the Audio Engineering Society, Vol 65(12), 1071-1031, (Dec 2017) [1].

Agus, N., Anderson, H., Chen, J.M., Herremans, D., Lui, S., "Perceptual evaluation of measures of spectral variance," Journal of the Acoustical Society of America, Vol 143(6), (Jun 2018)

Agus, N., Anderson, H., Chen, J.M., Lui, S., "Minimally Simple Binaural Room Modelling Using a Single Feedback Delay Network," Journal of the Audio Engineering Society (Jun 2018). Manuscript accepted (with editor).

[1](*) Both authors contributed equally

Acknowledgements

I would like to thank my supervisors Dr. Simon Lui and Dr. Dorien Herremans for supporting me throughout my PhD studies. I am very grateful that they have given me the freedom to work on a research topic of my choice and that they were always there to guide me along the way. They have also given me many opportunities to pursue my dream, which is teaching, while I was completing my PhD journey. Thank you very much for having so much faith in me since the very beginning.

I would also like to also express my utmost gratitude to my co-authors, Dr. Hans Anderson and Dr. Chen Jer-Ming for their valuable inputs and all the countless hours that we have spent together for this research. I would have never imagined that a small idea that we had three years ago would have grown so much to be our dissertation topic. Thank you for every single interesting discussions we had over the years. I have been very fortunate to have the opportunity to work with them.

A very big thank you also goes to my Mama and Papa, and Mollie. I am so blessed to have them at home and I cannot thank them enough for never stop inspiring and encouraging me to pursue my dreams, no matter how difficult things may get. I hope I have made them proud. My good friends, Joleen, Carmen, Crystal, Huiying, Vu, and Minh, thank you for always being my emotional support, especially during rougher days.

Finally, I would like to sincerely thank the rest of the thesis committee members, Prof. Zhou Jianying and Dr. Hyowon Lee for their time and valuable inputs.

Contents

List of Figures

List of Tables

For Papa and Mama…

Chapter 1

Binaural Auralization

Parts of this chapter is based on the technical report: Agus, N., Anderson, H., Chen, J.M., Lui, S., "Energy-based binaural acoustic model- ing," Singapore University of Technology and Design, Tech Report No. 1, (Apr 2017).

1.1 INTRODUCTION

In the literature, the methods for auralization can be generally classified into three different types of method: convolution, numerical, and geometrical acoustics. Each method has different pros and cons, briefly summarized as follows. Convolution methods using pre-recorded (binaural) room impulse responses (RIR) are accurate but inflexible, meaning that one recorded RIR specifically reproduces one source-listener configuration in a fixed room, and it cannot be adjusted to simulate different configurations or rooms. Numerical methods are also accurate and more flexible than convolution methods, but require extensive computational power. Geometrical methods are flexible and fast, but they compromise on accuracy. Therefore each method is suited for different types of applications.

Convolution of an input signal with a pre-recorded RIR embeds the acoustic response of that room into the input signal, as if it was played from and listened to at the source and listener locations of which the RIR was recorded respectively. It seems to be the easiest method to auralize a particular input signal, i.e: simulating how this new signal sounds like in a particular room of space, but this method is not flexible. The most obvious problem to this method is that one RIR is required for each source-listener configurations in the same room, leading to infinitely many combinations if the source or listener are allowed to move freely within the room (as is the case with many virtual reality systems). In some applications such as preliminary design of acoustic spaces like concert halls, the rooms in question are not physically available, thus this method cannot be used to study its acoustical properties before it was actually built. Convolution method is however useful in film and music production and it was popularized in early 1990s to add subtle reverberation effects that give the impression of space on audio signals (Välimäki et al., 2012).

In numerical acoustics, we can simulate room acoustics by numerically solving the wave equation using finite element method (FEM), finite-difference time-domain method (FDTD), or any other numerical analysis methods. Numerical acoustics is physically accurate, and is able to capture wave-based phenomena such as diffraction and interference. However, computation cannot be done in real-time and it requires a lot of computational resources. In (Raghuvanshi, Narain, and Lin, 2009), it is noted

that precomputation time of 15 hours was required to render sound for up to 1Khz in a large cathedral scene (35m x 26m x 15m). Another work (Raghuvanshi et al., 2010) that introduces novel ways to simplify the numerical methods still require approximately 2 hours of precomputation time to render sound (up to 1Khz) in a smaller size room (19m x 19m x 8m). Both computations in (Raghuvanshi, Narain, and Lin, 2009) and (Raghuvanshi et al., 2010) were done using quad-core 2.8Ghz Intel Xeon CPU. Even though nowadays there exist processors with faster clock-speed, it is still computationally impossible to solve the wave equation across all octave bands for the entire duration of impulse responses (IRs) in larger rooms (Välimäki et al., 2012), which can mostly range between 0.5s to more than 2s long in classrooms and large halls. Hence, numerical methods are ideal for in-depth study of acoustic spaces in limited frequency bands, such as investigating the acoustic behavior of concert hall designs where fast computation is not a priority.

The third class of methods is called geometrical acoustics (also called ray-based acoustics), which is much faster to compute than numerical methods as its fundamentals lie on the assumption that sound waves behave like light waves. Popular methods in geometrical acoustics include image source method (Allen and Berkley, 1979) and ray tracing. The main weakness of this method is that it ignores the effects of diffraction and interference. Diffraction and interference mainly occurs in the lower frequency bands, where sound wavelength is comparable to the size of everyday objects (>1m). Therefore on average, geometrical methods are most inaccurate when modeling frequencies below 300Hz (Siltanen, Lokki, and Savioja, 2010). This however may not pose much of a problem as on average, humans are not very sensitive in hearing such low frequencies. Although human can generally hear frequencies between 20-20kHz, studies have shown that human hearing is most sensitive only in the midband frequencies between 2-7kHz (Shaw and Teranishi, 1968). This is mainly attributed to resonances in the ear canal and concha that boost the amplitude of frequencies in this range. In (Ballachanda, 1997), it is stated that the pinnae also reduces the presence of low frequencies. Therefore despite being inaccurate in modeling wave phenomenon in the low frequency bands, geometrical methods are still able produce perceptually convincing outputs. Ideal applications of geometrical methods include commercial gaming and virtual reality systems for entertainment that require real-time processing at a reasonable cost, implying the need for lower computational requirement.

Since each of the methods above have their own benefits and drawbacks, many studies came up with hybrid methods which combine the best parts of the original algorithms used. For example, the work in (Murphy et al., 2008) used numerical acoustics to model wave phenomena only in the low frequency bands, and used geometrical acoustics method to model the rest of the frequency bands. The work in (Vorländer, 1989) utilized two different geometrical acoustics methods, ISM and ray-tracing, to model the early reflections (approximately the first 80ms of the IR) and late reverberation part (after 80ms) of the IR respectively. Authors in (Wendt, Par, and Ewert, 2014a) proposed a design that compute only the early reflections part of the IR using ISM, and only approximate the reverberation tail using an artificial reverberator, without explicitly computing each and every reflection. This is possible because since the number of reflections in this late part of IR is too dense, the reverberation tail generally resembles an exponentially decaying random noise. One may approximate the entire

late reverberation tail separately using a much simpler mathematical model (Lehmann and Johansson, 2010) or using an artificial reverberator like the Feedback Delay Network (FDN) (Jot and Chaigne, 1991). The FDN proposed by Jot and Chaigne in (Jot and Chaigne, 1991) remains as the most efficient and widely used artificial reverberator today (Välimäki et al., 2012).

1.2 MOTIVATION

We notice that many of these room simulation algorithms in the literature (some include (Wendt, Par, and Ewert, 2014a; Sena et al., 2015; Raghuvanshi et al., 2010; Bai, Richard, and Daudet, 2015)) are unable to perform computation in real time or auralize input signals algorithmically. To auralize input signals, they first produce a RIR based on the room parameters. Then secondly, convolution between input signals and resulting RIR is done. While this second step can be done in a faster way using the overlap-add method [1], it still may take away considerable computational resources especially when the required refresh rate is high. The cost is more apparent where resources is limited, such as in mobile devices.

For many virtual reality applications, it is important to render realistic graphics as much as possible. A normal person may easily tell the difference between 1080p and 4K video or image resolution. This is much attributed to the fact that the resolution of an average human eye is more than 500 megapixels. Most of the computational resources like the graphic processing unit should be dedicated to render graphics. Human auditory perception however does not work in the same way as the visual perception. Audio rendering does not require the same level of detail as graphics rendering to achieve perceptual plausibility. This is because of the fact that human tends to blend auditory image when there is not enough time delay between arriving sound wavefronts, and the perceived location of the source is heavily influenced only by the first wavefront. This effect is also known as the precendence effect (Wallach, Newman, and Rosenzweig, 1949) and Haas effect (Haas, 1972). Such effect is stated to remain valid even if the later reflections are louder by as much as 10dB, as long as these later reflections are within 25-35 ms of the first arriving reflections. It shows that human auditory system ignores or do not process much information from these later reflections.

In this light, our research focuses on developing a minimally efficient binaural room simulation algorithm design that it is able to directly auralize input signals in real time without the need of convolution, and yet remains perceptually convincing. We propose an algorithm that is able to auralize dry input signals in real-time, even on mobile devices, with acceptable degree of perceptual plausibility. Our research falls under geometrical acoustics category, which assumes that sound rays behave exactly like light rays and hence methods from computer graphics can be borrowed.

[1] convolution with long signals can be computed as a sum of many short convolutions, which is faster than performing traditional convolution with the entire signal length.

1.3 CHAPTERS OVERVIEW

To explain our proposed method clearly, the rest of this work is arranged as follows,

1. **Basic theories and mathematical framework**:
   In the rest of this chapter, we are going to summarize basic theories that are essential for understanding the proposed method. They are the Acoustic Rendering Equation (Siltanen et al., 2007), the Feedback Delay Network (Jot and Chaigne, 1991), and basic concepts of room impulse response.

2. **Chapter 2: Perceptual evaluation of measures of spectral variance**:
   An ideal Feedback Delay Network without attenuation gain should produce an output that resembles white noise, if one were to feed an impulse into the network. We use the Feedback Delay Network in our proposed system, and we set the delay lines based on first order reflection paths. We need to first check if such setting does not add too much coloration to the network. However, there was no work in the literature that established the just noticeable difference for sound whiteness, i.e: to answer the question, how much *coloration* can we add to an ideal white noise such that the noise is no longer perceived as being white. Therefore in this next chapter, we present our work where we established the Just Noticeable Difference for *whiteness*. We then present the study that our method does not noticeably alter the *whiteness* of the network output.

3. **Chapter 3: Whiteness of Lossless FDN**:
   We have established two things so far at this point. Firstly, the physical significance of signals inside the FDN and secondly, the JND of perceptual variance. By adding a physical significance of signals inside the FDN, we would need to set the FDN delay lengths based on the first order reflection paths from the source, to one of the surface geometries in the room and finally to the listener. We call this *ray-tracing delay lines*. This is fundamental for our real-time binaural auralization algorithm presented in the next chapter. However, before that we would like to first investigate on whether these ray-tracing delay lines will bring about unwanted artifacts to the FDN output. Ideally, the output of a lossless FDN (or any artificial reverberator) when an impulse is fed is desired to be as white as possible, meaning that the FDN does not alter the frequency component of the input in any way (Schroeder, 1962). However the output of the FDN is highly dependent on its delay lengths setting. In this chapter, we explained how the output of FDN is affected by the combination of its delay lengths, and present the result that ray-tracing delay lines do not colour the output of this FDN beyond that of noticeable amount (JND).

4. **Chapter 4: Minimally simple auralization algorithm**:
   In this Chapter 4, we present our proposed model, which is the full design of our proposed method. Our proposed method is a hybrid algorithm, where we explicitly model only first order reflections and approximate the rest of the reflection orders using a Feedback Delay Network. We balance the energy between the first order and the higher orders of reflection, and we show that the network does not contain any perceptually noticeable coloration using the Just Noticeable Difference we studied in Chapter 2.

5. **Chapter 5: Balancing the energy between early and late reflections in hybrid models**:
   In the following Chapter 5, we present the second part of our proposed model, which show how we can also apply our concept of balancing the early reflections and late reflections energy in existing hybrid auralization algorithms (methods that compute early reflections explicitly but only estimate the late reverberation for faster computation time). We identified the cause of energy balancing issue that is apparent in existing hybrid models, and offered a solution that can be integrated in existing hybrid models.

6. **Chapter 6: Adaptive room decomposition for auralization**:
   In this Chapter 6, we present the third and final part of our proposed model, which is a new method to subdivide the room surfaces for Monte-Carlo computation using The Acoustic Rendering Equation. In Chapter 4, we subdivide the room surfaces geometry evenly. The listening test result in this chapter shows that by subdividing the room polygons using this new method, perceptual cues pertaining to localization is greatly improved.

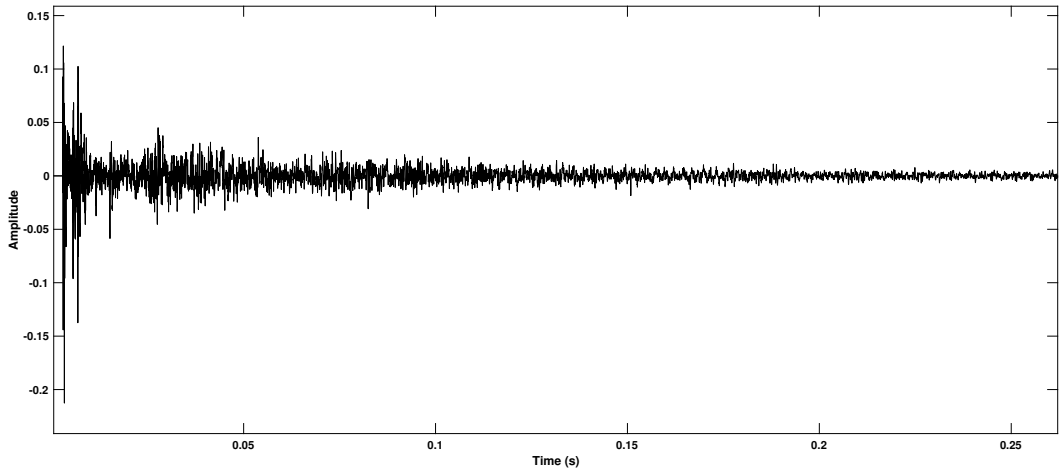

FIGURE 1.1: A sample impulse response.

1.4 ROOM IMPULSE RESPONSE

This section gives a brief overview of the parts of RIR that is relevant to this work, especially the subjective qualities of room impulse response (RIR). We refer readers to (Schroeder et al., 2007) and (Kuttruff, 2009) for in-depth details on room acoustics.

Figure 1.1 shows a sample plot of a real recorded IR. The x-axis indicates the time axis in seconds, and the y-axis indicates its amplitude, which translates to the loudness of the signal. There are several ways to obtain such IR. The simplest way can be done using Dirac (an ideal impulse), either from popping a balloon or firing a blank pistol (Kuttruff, 2009). The idea is to produce all audible frequency range in a split second and observe how they decay over time. These methods however are known to be unstable since they may not cover the entire frequency range and are also affected by background noise. A more robust way to measure an RIR is by using sine sweep method (Farina, 2000). We will elaborate in further detail how we obtain and record RIRs in Chapter 4.

Typically, we can divide the RIR into three parts (Funkhouser, Jot, and Tsingos, 2002; Griesinger, 1999; Litovsky et al., 1999). The first part is the direct sound, which translates to the first sample that makes up the entire impulse response. This resembles the sound waves that hits the microphone or listener without being reflected off any surfaces first. The second part is called the early reflections, which is the first 50 to $80ms$ of the RIR (Hidaka, Yamada, and Nakagawa, 2007). Typically this includes the first to third order reflections (sound waves being reflected off surfaces once to three times before reaching the microphone or listener). The early reflections are usually seen as the part of the RIR that carry the most perceptual cues and spatial information because on average, human ears are able to distinguish the individual reflections (Kuttruff, 2009; Välimäki et al., 2012; Griesinger, 2010). The third part is called the late reverberation, which is the samples of the IR from $80ms$ onward. The samples in the late reverberation part of the IR is densely clustered and evenly mixed. It is not perceptually possible for humans to distinguish each reflections from the late reverberation part of the IR. We

can generalize this part of the IR as a stochastic process with exponentially decaying amplitude envelope. This part of the IR conveys the sense of *spaciousness* in the room (Griesinger, 1996). In the literature, this is called the *reverberation time* (abbreviated as RT_{60}), which is the time taken for the impulse to decay for 60dB. A room with longer RT_{60} is perceived as a bigger room and vice versa.

It is known to be computationally infeasible to explicitly compute every single sample in real time for the entire duration of the impulse response, even with the current state of the art technology (Välimäki et al., 2012). Authors in (Antani and Manocha, 2013) even claim that it is nearly impossible to accurately model individual reflections beyond 4th order reflections and above. Therefore for real-time applications, the accuracy of the late reverberation part of the IR is often reduced into an approximation. There are many ways to estimate the reverberation tail, either using statistical model (Schroder and Vorlander, 2007) or artificial reverberators such as the FDN (Jot and Chaigne, 1991).

1.5 EXISTING TECHNICAL BACKGROUND

In this section we briefly explain two existing works of which this research relies upon. The first work is the ARE (Siltanen et al., 2007), and the second work is the FDN (Jot and Chaigne, 1991).

1.5.1 The Acoustic Rendering Equation

In computer graphics, the ongoing challenge for producing realistic scene is to come up with methods that can approximate and solve the rendering equation (Kajiya, 1986). In 2007, Siltanen et. al introduced the Acoustic Rendering Equation (ARE) (Siltanen et al., 2007), which is the acoustic counterpart of the rendering equation in (Kajiya, 1986). The ARE serves as a unifying framework for all geometrical acoustic algorithms, since all of them can be described using the ARE (Välimäki et al., 2012). In other words, the ARE (Siltanen et al., 2007) is an integral equation that governs the the behavior of all geometrical acoustic modeling methods. The ARE is expressed as follows,

$$\ell\left(\boldsymbol{x},\Omega\right)=\ell_0\left(\boldsymbol{x},\Omega\right)+\int_{\mathcal{G}}R\left(\Lambda_{[\boldsymbol{u},\boldsymbol{x}]},\boldsymbol{x},\Omega\right)\ell\left(\boldsymbol{u},\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}\right)\mathrm{d}\boldsymbol{u}. \tag{1.1}$$

The ARE can be seen as the audio counterpart of the Kajiya rendering equation used in computer graphics (Kajiya, 1986). The term $\ell(\boldsymbol{x},\Omega)$ represents radiance from a surface point $\boldsymbol{x}$ to a specific direction Ω. Radiance from a surface point is made up of two components. First, the left hand term in Equation 5.4, ℓ_0, represents emitted radiance from that point if $\boldsymbol{x}$ is a sound source. The second term in Equation 5.4 represents the sum of other input acoustic radiance from the rest of the room surface $\mathcal{G}$. R is known as the reflection kernel. It determines how much of the radiance coming from the point $\boldsymbol{u}$ is reflected off $\boldsymbol{x}$ to the direction Ω. Graphically, Equation 5.4 is represented by Figure 1.2.

To simplify notation in the following sections, we define $\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}$ to be a unit vector pointing in the direction from $\boldsymbol{u}$ to $\boldsymbol{x}$,

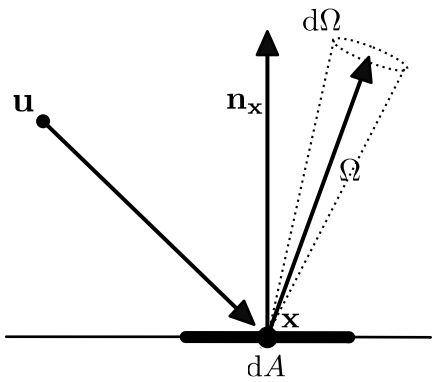

FIGURE 1.2: A ray from point $\boldsymbol{u}$ is reflected off surface point $\boldsymbol{x}$.

$$\Lambda_{[\boldsymbol{u},\boldsymbol{x}]} = \frac{\boldsymbol{x}-\boldsymbol{u}}{\|\boldsymbol{x}-\boldsymbol{u}\|}. \tag{1.2}$$

Therefore, Equation 5.4 can be rewritten into,

$$\ell\left(\boldsymbol{x},\Omega\right) = \ell_0\left(\boldsymbol{x},\Omega\right) + \int_{\mathcal{G}} R\left(\Lambda_{[\boldsymbol{u},\boldsymbol{x}]},\boldsymbol{x},\Omega\right)\ell\left(\boldsymbol{u},\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}\right)\mathrm{d}\boldsymbol{u}. \tag{1.3}$$

The reflection kernel R is consisted of an absorption function ξ (attenuation due to propagation losses in air.), a visibility function $\mathcal{V}$, a bidirectional reflection distribution function (BRDF) ρ, and a geometry function g,

$$R(\Lambda_{[\boldsymbol{u},\boldsymbol{x}]},\boldsymbol{x},\Omega) = \xi(\boldsymbol{u},\boldsymbol{x})\ \mathcal{V}(\boldsymbol{u},\boldsymbol{x})\ \rho\left(\boldsymbol{u},\boldsymbol{x},\Omega\right)\ g(\boldsymbol{u},\boldsymbol{x}). \tag{1.4}$$

The formula for ξ in linear absorptive medium is (Siltanen et al., 2007),

$$\xi(\boldsymbol{u},\boldsymbol{x}) = e^{(-\varepsilon\|\boldsymbol{u}-\boldsymbol{x}\|)}. \tag{1.5}$$

$\mathcal{V}$ takes a value of 1 if $\boldsymbol{u}$ is visible from $\boldsymbol{x}$ and is 0 otherwise. g models the effect of the distance (inverse-square distance law (Schroeder et al., 2007)) between $\boldsymbol{u}$ and $\boldsymbol{x}$ and the orientation of the respective surface normals : $\boldsymbol{n_u}$ and $\boldsymbol{n_x}$ on the amount of energy propagation between the two arbitrary surface points,

$$g(\boldsymbol{u},\boldsymbol{x}) = \frac{\left(\boldsymbol{n_u}\cdot\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}\right)\left(\boldsymbol{n_x}\cdot\Lambda_{[\boldsymbol{x},\boldsymbol{u}]}\right)}{\|\boldsymbol{u}-\boldsymbol{x}\|^2}. \tag{1.6}$$

In (Siltanen et al., 2007), the author includes a time delay and absorption operator in the geometry term. However in this work, we take the absorption operator outside of the geometry term and omit the time delay operator. We found that by rearranging the formula in this way, we are able to present our model with more clarity and consistency.

The surface normal $\boldsymbol{n_u}$ for a point source is undefined. But we can define the surface normal at the point source $\boldsymbol{n_u}$ to be equal to $\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}$. When we define the surface normal for point source in this way, the first dot product in the numerator of (1.6) is always equal to 1,

$$\boldsymbol{n_u} \cdot \Lambda_{[\boldsymbol{u},\boldsymbol{x}]} = 1. \tag{1.7}$$

The BRDF in (1.4) is denoted by the symbol $\rho(\boldsymbol{u}, \boldsymbol{x}, \Omega)$. It contains the information on the reflective properties of the surface (diffuse and specular content), i.e: how much radiance is reflected from the $\boldsymbol{u}$ to direction Ω from $\boldsymbol{x}$. We can estimate BRDF mathematically up to a certain amount of accuracy (Kiminki, 2005), or measure it using a physical sample.

All geometrical acoustic methods attempt to solve or approximate the ARE efficiently using various methods. The complexity of solving the ARE grows exponentially with respect to the order of reflections since sound rays that hit any surface can be scattered to all directions. Hence most methods only solve the ARE up to the third order (Välimäki et al., 2012; Välimäki et al., 2016), and approximate the reverberation tail using artificial reverberator or statistical model. One can prevent the exponential growth in complexity by computing only specular reflections. The image source method (Allen and Berkley, 1979) does exactly this. Each surface strictly reflects incoming sound ray to one specific direction, depending on the angle of incident, instead of scattering it in all directions.

1.5.2 Feedback Delay Network

The FDN (Jot and Chaigne, 1991) is a popular and efficient artificial reverberator. The generic structure of FDN is shown in Figure 1.3. μ and υ represents (optional) input and output gain respectively. g represents exponential decay gain that can be applied to control the reverberation time. The mixing matrix used in an FDN has to be unitary, meaning that if the value of g is set to be 1, the system is lossless. In other words, if we were to input a single impulse to the system, we will get infinite output. The gain g can be set according to the following formula,

$$g_j = 10^{3*d_j/RT_{60}} \tag{1.8}$$

so that the system will decay according to the desired reverberation [2] time RT_{60}. A standard setting for input gain μ is,

$$\mu_j = \frac{1}{\sqrt{N}} \tag{1.9}$$

Setting μ to the value in Equation 1.9 makes input energy equivalent to the total output energy when the signal has passed through all the delay lines for the first time. Perceptually, it means that it preserves 'loudness'.

[2] RT_{60} is the time taken for sound pressure to decay by 60 dB after source has been turned off.

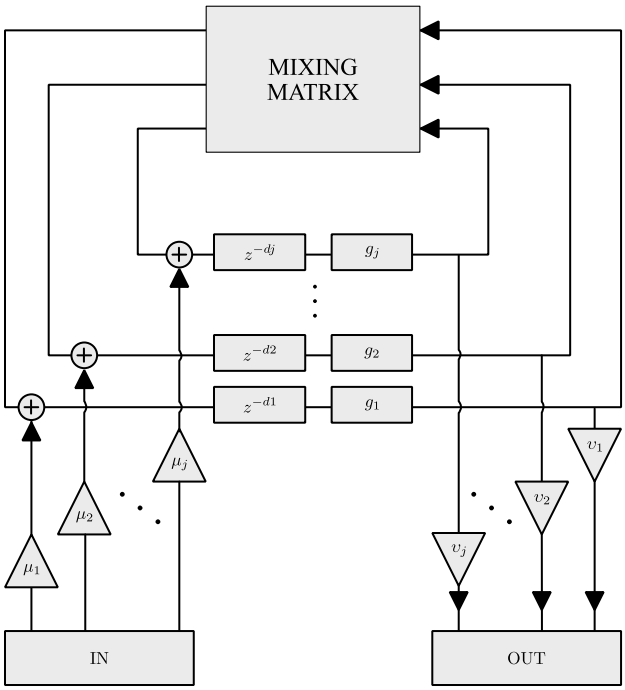

FIGURE 1.3: Basic FDN structure with N delay lines proposed by Jot and Chaigne in 1991.

1.6 MATHEMATICAL PRELIMINARIES

This section explains basic mathematical framework that is essential for understanding our proposed method in the rest of the chapters. Parts of this section is based on our technical report in (Agus et al., 2017). Some of these concepts are borrowed from radiometry, and we refer readers who are interested in more in-depth explanations to (McCluney, 2014). We wrote this technical report because we found that there is lack of theoretical explanation in the literature that bridge the gap between understanding radiosity theory in (Kajiya, 1986) and acoustic theory in (Siltanen et al., 2007). The report contains theoretical foundation that explains how techniques from radiosity can be applied to compute acoustic energy. We summarize parts of this report that is relevant to this work in this section.

1.6.1 Basic Mathematical Framework

Sound waves that propagate through the air cause deviations in the local air pressure from atmospheric pressure. Digital audio signal represents microphone measurements of these air pressure deviations. The ARE is an acoustic energy propagation model, hence we need to convert the physical meaning of the input signal in terms of energy before using it in the ARE. We do this firstly by representing sound pressure in terms of acoustic intensity, and then using the relationships between acoustic radiance and energy flux to come up with a quantity for the input signal's total energy flux.

In the literature, it is widely known that (instantaneous) acoustic intensity [3] of plane wave, $\boldsymbol{I_a}$, is a vector proportional to the square of its pressure (Schroeder et al., 2007),

$$\boldsymbol{I_a} = \boldsymbol{r}\frac{p^2}{\rho c} \tag{1.10}$$

ρ and c represents density of the medium of propagation and speed of sound in that medium respectively. The vector $\boldsymbol{r}$ represents the unit vector pointing into the direction of the acoustic energy flux. Note that acoustic intensity is time variant since p, the sound pressure, varies with time. A more precise notation is $p(t)$ for pressure and $\boldsymbol{I_a}(t)$ for acoustic intensity, but to simplify our notations we omit that in our explanation.

There exist a more commonly mentioned concept, that *sound pressure is inversely proportional to distance* or equivalently *acoustic intensity is inversely proportional to the square of distance*. Mathematically, we often see the following expression,

$$p_i \propto \frac{1}{d_i} \tag{1.11}$$

In geometrical acoustics, we would like to compute p_i, where i will be the listener's location. So given p, the input signal, a very simple way to compute pressure at location i (set to be d_i distance away from the source) will have a pressure of $p_i = p\frac{1}{d_i}$.

In a room, we can imagine when a source emits sound waves, they hit objects and walls present in the room, and then these surfaces reflect (some of) these waves back to the listener. Any surface in the room can be seen as a secondary sound source. Using the ARE we can compute how much energy each surface contains (or equivalently the sound pressure at each surface point), which in turn allow us to compute the sound pressure at listener's location.

We can immediately see two problems that may rise from Equation 1.11 when $d_i < 1$. Firstly, digital output signals are limited between range $-1 to 1$. As $d_M \to 0$, $p \to \infty$ and clipping might occur, resulting in inaccurate relative energy modeling between sound rays.

Secondly, if a listener is placed very near a particular surface, this surface may sound particularly loud, even possible to be louder than the source itself from some distance and dominate the rest of the rays. For example, we set the distance between the listener and a wall to be 5 cm, and the distance between that wall to an omnidirectional sound source to be 1m. Assuming that the wall diffusely reflects 90% the sound energy, we would hear an output of $p/1 \times 1/\pi \times 0.9/0.05 = 5.3p$ from that single wall reflection. This is 5.3 times louder than the source at 1m distance. Such phenomenon may potentially render unrealistic output. It is generally counterintuitive to hear that a patch of surface can potentially reflect sound that is louder than the source itself.

This points out that we need to establish some kind of minimum distance d_M to prevent Equation 1.11 to reach infinity and to also ensure conservation of energy (preserving the correct relative amount energy between rays) in the computation of ARE. Conservation of energy can be guaranteed when we establish d_M as we can keep track

[3] Instantaneous net flow of sound energy (flux) emitted, transmitted, reflected, or received through (any, not necessarily physical) unit area in the direction perpendicular to the unit area. Time-averaged acoustic intensity is the root-mean-square of $\boldsymbol{I_a}$. The dot product of $\boldsymbol{I_a}$ with a physical surface normal yields what we know as sound power (flux) through that particular surface.

of the maximum amount of energy in the system. This will be explained later in this section (Equation 1.14).

Intuitively d_M can be set as *the minimum distance that a listener can get close to any subject*. Physically, this distance is the length of the ear canals, assuming the source is outside of the listener's body. The closest a listener can get to a speaker in the room is when the listener places his or her ear directly at the speaker's grill. This scenario will be the maximum (loudest) amount of flux that the listener can collect and in this way we will never encounter such scenario where a surface point can act as a secondary source that is louder than the input signal itself.

Once we have set d_M, we can compute p_i as,

$$p_i = p\,\frac{d_M}{d_i},\ d_i > d_M, \tag{1.12}$$

The distance between a listener and any surface point in the room cannot be lesser than d_M to keep the relative energy between surface points in the room consistent during calculations using the ARE. Larger d_M makes the overall output sound louder than smaller d_M.

The total instantaneous flux energy that passes through an enclosed surface $\mathcal{G}$ around a source with acoustic intensity I_a is defined as the surface integral,

$$\Phi(\mathcal{G}) = \int_{\mathcal{G}} \boldsymbol{I_a} \cdot \boldsymbol{n}_A \,\mathrm{d}A. \tag{1.13}$$

If we assume that $\mathcal{G}$ is a spherical enclosure around the source with a certain radius d_M, and the source emits energy radially, such that $\boldsymbol{r}$ is always pointing to any receiving end (listener or any surface in the room), we have,

$$\Phi(\mathcal{G}) = 4\pi d_M^2 \frac{p^2}{\rho c}. \tag{1.14}$$

Here it also becomes clear that d_M has to be nonzero in order for Φ to exist (> 0). As $d_M \to 0$, $Phi \to \infty$. As mentioned, a physical meaning to this minimum distance is the distance between the source and the eardrum. In reality, d_M will never be 0, as a point source is not physically feasible. Any sound source needs to have a form of physical shape, or mass, e.g: a speaker, a musical instrument, objects clanging together, etc, and an ear canal is definitely non-zero.

In summary, establishing d_M sets the total source flux using Equation 1.14, and this will be useful to ensure conservation of energy when we compute acoustic energy flux on various surfaces in the room using the ARE.

1.6.2 Physical Significance of Audio Signals in FDN

The FDN can be used to add artificial reverberation on any input signal. We only need to set its size, reveberation time, and delay lengths to achieve the desired effect. It can certainly be used in room simulation algorithms, some of the many examples include (Jot, 1997; Menzer, 2012; Menzer, 2012; Savioja et al., 1999) to produce good sounding reverberation tail of the RIR efficiently. Unlike the ARE, it doesn't explicitly compute the energy or sound pressure for each and individual reflections based on physical

properties, and there is no (prior) physical meaning of signals inside in the FDN when it is used in room simulations. In other words, FDN is commonly used to produce the late reverberation tail of an RIR in room modeling systems not because it is or mathematically accurate [4] but because it is capable to produce exponentially decaying sound with high echo density that sounds very similar to a real RIR's reverberation tail.

The FDN is consisted of multiple delay lines and a mixing matrix (see Figure 1.3). As these signals circulates in the FDN over time, the original signals from each delay lines will be evenly mixed [5], meaning that each output tap of the FDN will eventually contain approximately equal amount of energy over time. This is what makes FDN a perfect candidate to approximate the late reverberation tail of an RIR, which is often characterized as having high echo density or diffused. An RIR is basically made up of sound reflections in the room which grows exponentially as the number of reflection order increases. Early order reflections (up to third order) typically contain considerable amount of specular reflections. The perception of diffuse reflections dominate in higher order reflections (Kuttruff, 2009). This is not because the reflection behavior of each surface point in the room changed over time, but because perceptually, humans are no longer able to distinguish individual reflections, therefore allowing us to model the late reflections as a diffuse process. Energy from diffuse reflections are much simpler to model in the ARE than specular reflections since we do not have to trace the path of each sound rays which complexity grows exponentially with respect to the number of reflection order. This statement will be apparent later on in Chapter 4.

Previously (Section 1.6.1), we explained that time-varying digital input signal p represents measurements of sound pressure at the microphone location. We attempt now to offer the physical significance of signals inside the FDN when it is used to model room acoustics. The point of doing this is that if we can establish physical significance of audio signals inside the FDN, we can use it to model earlier reflections and not just the late reverberation tail efficiently. For example, in (Wendt, Par, and Ewert, 2014a), FDN is used to approximate fourth and higher order reflections. Later on in Chapter 4 we propose room modeling algorithm that uses FDN to model second and higher order reflections, hence making it able to run in real time and auralize input signal algorithmically (directly) without the need for convolution. We propose that the following represents physical representation of audio signals in FDN,

1. The delay lines in the FDN can be set as the length for each ray in first order reflections, i.e: the time taken for sound to travel from the source to one of the surface points in the room, and then finally to the listener.

2. Hence, when signals exit the delay lines for the first time and meet the output taps, they can be seen as first order reflections. If appropriate amount of gains μ shown in Figure 1.3 are applied, we can exactly compute the amount of energy of each ray in first order reflections since each delay line represents one ray. In (Agus et al., 2017), we show direct application of ARE to compute these gains.

[4] The only ways the compute mathematically accurate RIRs for its entire duration is by solving the wave equation (all frequencies), or by solving the ARE (for higher frequencies).

[5] This is true if we chose an appropriate unitary feedback matrix that optimize mixing and scattering, such as the Hadamard matrix (Smith, 2010).

3. These signals will also be multiplied by the mixing matrix. This process can be thought as sound waves mixing as they hit the surfaces in the room for the second time, since the matrix's primary function is to scatter and mix signals in the delay line.

4. When signals exit the delay lines for the second time to the output taps, they can be seen as 'second order reflections'. However unlike first order reflections, this is just an approximate value. The reason to this is obvious. Firstly, we do not have the correct amount of rays since the amount of rays in second order reflections will be squared of the amount of rays in first order reflections. Secondly, we do not have the correct length delays for each individual second order reflections. Thirdly, we do not set any new gain values that resembles the amount of energy received in listener location from second order reflections and therefore the amount of energy in the 'second order reflections' is not accurate.

5. When signals exit the delay lines for successive times, they can be seen as 'third', 'fourth', and successive order reflections respectively. Again, the output is only an approximation and it is not mathematically accurate.

6. As signals circulate in the FDN for more iteration, the signals entering the output taps become more diffused. In other words, each output tap carries approximately the same amount of acoustic energy. This physically resembles higher order reflections, where signals in the room have been evenly mixed such that on average, each surface point can be seen as a diffused reflecting medium. Therefore the accuracy of the FDN in modeling higher order reflections increases (e.g: the accuracy in modeling sixth order reflection is higher than second order reflection), provided that we set υ accordingly using the ARE (Agus et al., 2017).

By adding physical meaning of audio signals inside the FDN, we can model interaural effects not only in early reflections but also in the late reflections. We theorize that it is therefore sufficient to perform computation for each of the first order reflections ray and significantly reduce the computation time [6]. We subsequently evaluate and support this claim in (Agus et al., 2017). The complexity of the model in (Agus et al., 2017) is linear with respect to the number of sound rays used. On contrary, if we were to compute higher order reflections (second order onwards), our computational complexity would have been exponential with respect to the number rays.

Since the signals that circulate inside the FDN is in terms of sound pressure, in order to successfully use the ARE to perform acoustic energy-based computations, we can think of these signals at the output taps of the FDN as the square-root of energy flux leaving all surfaces in the room. Mathematically, the total energy flux from the FDN (before multiplication by output gain υ) is,

$$\Phi(\text{FDN}) = \sum_{i=1} N y_i^2, \tag{1.15}$$

where y_i is the output of the ith delay in the FDN. If the FDN is lossless, (i.e: g_i is set to 0), and if μ is set to be energy preserving as Equation 1.9, then according to conservation

[6] when compared to other acoustic room simulation designs that perform computations for second and third order reflections.

of energy, $\Phi(\text{FDN}) = 4\pi d_M^2 \frac{p^2}{\rho c}$, the total (instantaneous) source energy quantity derived earlier in Equation 1.14.

1.6.3 Setting FDN delay lines

By assuming that it is possible to establish physical significance of audio signal in the FDN as explain in the previous Section 1.6.2, we can set the length of FDN delay lines as the time taken for sound waves to propagate to the listener during the first order reflection. However, the setting of the delay lengths in FDN will affect its output coloration. Ideally, a FDN should produce an output that resembles white noise when an impulse is fed through it and when the decay is set to zero (Smith, 2010). This simply means that the FDN does not alter the color of the input signal, which is a property of audio filter that is desirable in practice (Rocchesso; and Smith, 1997; Rochesso, 2000; Rocchesso, 1997; Dahl and Jot, 2000).

Therefore the first step towards designing our binaural auralization model is to investigate whether setting the delay lengths of the FDN using the first order reflections path introduces too much coloration. From this point onwards, we refer to this delay line setting as as *ray tracing delay lines*. The problem with this is that there has been no study in the literature that investigates the minimum amount of coloration that can be added to a an ideal white noise such that the noise is no longer perceived as white. Therefore in the next chapter, we present our work that establish the Just Noticeable Difference of *spectral variance*, which is the noticeable amount of color that can be present in a white noise before subjects perceive the noise as tonal (coloured). We then use the Just Noticeable Value found in Chapter 2 to evaluate the effect of ray-tracing delay lines on the output of a lossless FDN when an impulse is fed. This result is shown in Chapter 3. In Chapter 4 we also elaborate in detail and show that such assumption about the physical significance of audio signal in the FDN is reasonable and that our model is able to produce a perceptually plausible binaural RIR.

1.7 SUMMARY

This chapter contains a brief summary of some of our contributions in (Agus et al., 2017). We presented basic mathematical framework that transforms digital input signals in terms of acoustic energy for computations using the ARE. We also establish the physical significance of audio signals in FDN. This assumptions have two implications. Firstly, on whether the FDN is still ideal after such assumption, meaning that it still can produce ideal white noise in its lossless form (zero attenuation). Secondly, on whether this assumption is able to produce perceptually good binaural RIRs. To address the first implication, we need to first establish the Just Noticeable Difference for noise whiteness, also what we call as spectral variance. We present our study for this Just Noticeable Difference of spectral variance in the next Chapter 2. The second implication will be addressed in Chapter 4 and 5, where we show that our proposed system is able to auralize input signals with good objective and subjective ratings.

Chapter 2

Perceptual evaluation of measures of spectral variance

This work is based on the published manuscript: Agus, N., Anderson, H., Chen, J.M., Herremans, D., Lui, S., "Perceptual evaluation of measures of spectral variance," Journal of the Acoustical Society of America, Vol 143(6), (Jun 2018)

2.1 ABSTRACT

In many applications it is desirable to achieve a signal that is as close as possible to ideal white noise. One example is in the design of artificial reverberator, whereby there is a need for its lossless prototype output from an impulse input to be perceptually white as much as possible. In the previous chapter we introduced the ray-tracing delay line setting, which is an integral part of our binaural auralization algorithm. We need to investigate whether setting the delay line lengths in this way is perceptually acceptable, however there was no prior study in the literature that investigates the threshold for spectral flatness. Therefore in this Chapter we studied and establish such threshold. The Ljung-Box test, the Drouiche test, and the Wiener Entropy, also called the Spectral Flatness Measure are three well-known methods for quantifying the similarity of a given signal to ideal white noise. We conducted listening tests to measure the threshold, also known as the Just Noticeable Difference (JND) on the perception of white noise. In other words, this is the JND between ideal Gaussian white noise and noise with a specified deviation from the flat spectrum. We report the JND values using one of these measures of whiteness, which is the Ljung-Box test. We also found considerable disagreement between the Ljung-Box test and the other two methods and we show that none of the methods is a significantly better predictor of listeners' perception of whiteness. This suggests a need for a whiteness test that is more closely correlated to human perception.

2.2 INTRODUCTION

Contexts where we need to quantify the deviation of a signal from ideal white noise include linear predictive coding, perceptual audio coding, and designing digital reverberation networks (Harma and Laine, 2001; Johnston, 1988; Jot and Chaigne, 1991; Smith, 2010). Due to design constraints, we usually can not achieve perfectly white output and are forced to make trade-offs in the optimisation process. For that reason it would be helpful to know not simply how much variance a spectrum has but also to know whether or not that level difference from ideal white noise is audible. Some of the common measures of whiteness include the Wiener Entropy or Spectral Flatness Measure (SFM) (Gray and Markel, 1974), Ljung-Box test (Ljung and Box, 1978), and Drouiche Test (Drouiche, 2000).

When building an feedback delay network (FDN) reverberator, we normally begin by making a lossless prototype reverberator whose impulse response should resemble random white noise (Smith, 2010). An FDN typically comprises a bank of between 8 to 16 delay lines (Jot and Chaigne, 1991). After adjusting delay times to achieve a spectrally well-balanced result from the lossless prototype, we then apply additional filters and decay coefficients to achieve the desired decay time and spectral envelope (Jot and Chaigne, 1991). Hence to ensure that these filters produce the intended effect, the sound of the unfiltered lossless FDN should resemble white noise as closely as possible (Rocchesso; and Smith, 1997; Rochesso, 2000; Rocchesso, 1997; Dahl and Jot, 2000).

In linear predictive coding, we optimize a linear model to account for as much of a signal as possible, allowing us to reduce the data rate by coding only the coefficients of the linear predictor and the residual signal. Since an ideal model would account for all but the random component of a stationary signal, the goal of the model optimization can be expressed in terms of the whiteness of the residual signal; the more closely the residual resembles white noise, the more perfect the model (Cox, 1966; Makhoul and Wolf, 1972; Gray and Markel, 1974).

We present listening test results in this work, comparing zero-mean Gaussian white noise signals with other noise signals of known spectral variance. From the listening test results in Section 2.4, we derived JND values that indicate the smallest audible deviation of spectral variance from zero-mean Gaussian noise. The JND is the smallest change in a given parameter that is audible in more than fifty percent of the trials of listening test experiments (Fechner, 1966).

There have been studies on the perception of other aspects of white noise in the literature, such as the sensitivity and perception of interrupted white noise and periodic white noise, but not the attempts to establish a JND for spectral variance in specific. Miller et. al (Miller and Taylor, 1948) investigated the perception of short bursts of white noise. Pollack (Pollack, 1969) presented another study where he found that the periodicity pitch of interrupted white noise is factual. In (Wicke and Houtsma, 1975), Wicke et. al found that the musical pitch of interrupted white noise is rather weak. Duifhuis (Duifhuis, 1973) studied the audibility of harmonics in a periodic white noise. The perceptual sensitivity in the changes of white noise intensity has also been studied (Miller, 1947). In this work, however, we present a study on the *perception* of white noise itself.

In this paper, the spectral variance of all the noise signals and the JND value is presented in terms of $\hat{Q}$-value calculated using the Ljung-Box test. In section 2.5 we discuss the lack of agreement between the Ljung-Box test, which is based on autocorrelation, and the Drouiche test, which are frequency domain based methods. In section 2.5.2 we discuss the perceptual implications of this disagreement and suggest a direction for future study to design a more perceptually relevant measure of whiteness.

2.3 BACKGROUND INFORMATION

2.3.1 Measures for whiteness

In this section we briefly summarize three widely-used methods for quantifying the similarity between a given sequence of audio samples and ideal white noise: the Wiener Entropy or SFM, the Ljung-box test (Ljung and Box, 1978) and the Drouiche test (Drouiche, 2000).

One of the simplest methods for quantifying the spectral variance of a sequence of audio samples is to take the ratio of the geometric mean and arithmetic mean of the power spectrum. This is known as the Wiener Entropy, also called the Spectral Flatness Measure, abbreviated by the letters SFM. The ratio of geometric mean to arithmetic mean was first applied to audio signals in the time domain by Cox (Cox, 1966). The earliest work we could find that used this method to compute spectral flatness in the frequency domain is (Makhoul and Wolf, 1972). However, many authors cite a later work by (Gray and Markel, 1974) as the source.

Let x be an array of time series samples of length N,

$$x = \{x_1, x_2, x_3, ..., x_N\}, \tag{2.1}$$

and let $X = \{X(0), X(1), X(2), ..., X(N)\}$ denote the z-transform of x that is computed on the unit circle. The power of X is the squared magnitude of X, denoted as $|X|^2$. Then the SFM of X is,

$$\Xi(X) = \frac{(\prod_{n=1}^{N} |X(n)|^2)^{\frac{1}{N}}}{\frac{1}{N}\sum_{n=1}^{N} |X(n)|^2}. \tag{2.2}$$

The value of Ξ lies in the interval $[0, 1]$. A signal with a completely flat spectrum will have $\Xi = 1$ and the value of Ξ decreases as the spectral variance increases. A pure tone, with non-zero magnitude spectrum at only one frequency, would have an SFM of zero.

(Madhu, 2009) notes that the SFM is problematic due to the fact that if $X(n) = 0$ for any one of the frequency bins n then $\Xi(X)$ will be zero regardless of the variance of the remaining portion of the spectrum. To mitigate this problem he modifies Equation (2.2) as follows,

$$\hat{X} = \frac{|X(n)|^2}{\sum_{n=1}^{N} |X(n)|^2} \tag{2.3}$$

$$\log_2(\mathcal{G}+1) = -\frac{1}{\log_2(N)} \sum_{n=1}^{N} \hat{X}(n) \log_2(\hat{X}(n)) \tag{2.4}$$

The measure $\mathcal{G}$ provides a meaningful result even if the magnitude spectrum contains some zero values.

We generated 10 million noise signals with a zero-mean Gaussian amplitude distribution. Their length is 4096 samples each. We found their average SFM value to be 0.56. Since an SFM of one corresponds to a flat spectrum, this indicates that these Gaussian white noise samples are not white in the sense that they do not have flat spectra (Dougherty, 2009). Therefore we need to reiterate the definition of white noise in stochastic terms. From this point onwards, we use the term *Gaussian white noise (GWN)*, and we precisely mean noise that is the output of a zero-mean Gaussian random process with variance of 1, for which the *expected value* of the spectrum is flat. However, it does not mean that individual observations from such a process ought to have flat spectra. This definition takes a form that resembles a hypothesis test. For example, given that we observed a particular value of $\Xi(X)$, what is the probability that the sequence x could be the output of a stationary, zero-mean random process?

To make a precise calculation of this likelihood, it would be necessary to derive the probability distribution function (PDF) of $\Xi(X)$ under the assumption that x is a white noise output from Gaussian random processes. Unfortunately, an exact formula for this probability distribution function is not known. For this reason, we prefer to use spectral variance measures for which the PDF is known.

Ljung and Box discovered a now widely-used portmanteau statistic that was originally intended to test for lack of fit in time series models by calculating the average autocorrelation in the residual (Ljung and Box, 1978). This test is based on the idea that a perfectly fitted linear model should account for 100% of the components of the time series data that are stationary in the frequency-domain, leaving only pure white noise in the residual signal. Therefore, by comparing the residual against ideal white noise, we get an idea of how well the linear model fits the data. This idea of testing the lack-of-fit of a linear model was originally applied in the context of linear predictive coding (Gray and Markel, 1974). However, in the context of audio signal processing, the Ljung-box test is applicable in any situation where we want to compare an audio signal to white noise, even if the signal we are testing is not actually the residual error of a linear prediction.

The Ljung-Box statistic $Q(x)$ quantifies the average normalised autocorrelation in the signal over lag times between 1 and m,

$$Q(x) = N(N+2) \sum_{k=1}^{m} \frac{r_k^2}{(N-k)}, \tag{2.5}$$

where r_k is the autocorrelation at a lag value of k samples,

$$r_k = \frac{\sum_{t=k+1}^{N} x_t x_{t-k}}{\sum_{t=1}^{N} x_t^2}, \ (k = 1, 2, ...). \tag{2.6}$$

Note that unlike the SFM, higher values of Q indicate a higher amount of spectral variance. If the input signal is made up of zero-mean, independent and identically distributed (i.i.d.) $N(0, \sigma^2)$ random deviates then $Q(x)$ follows the Chi-squared (χ^2) distribution with m degrees of freedom, which has an expected value of m and variance of $2m$. If we define white noise to be the output of an i.i.d. random process, then it is impossible to say with absolute certainty that a given signal is or is not white noise, because for any signal we observe, there is a finite probability that a random process could generate such a signal. The advantage of using the statistic $Q(x)$, rather than the SFM mentioned previously, is that its probability distribution function is known. Therefore, for any signal x we can calculate a p-value that represents the probability of an i.i.d. random process generating a signal with the same or greater auto-correlation than x. In this way we obtain a statistically meaningful answer to the question of whether x is white noise or not.

The fact that the Ljung-Box method is based on auto-correlation of the signal in the time domain has perceptual implications that we will discuss in the next section. (Drouiche, 2000) proposed a method that allows us to make similar statistical statements about the whiteness of x but it is based on characteristics of its frequency-domain representation. The spectral density of x at frequency ω is defined by,

$$I_N(\omega) = \frac{1}{2\pi N} \left| \sum_{k=1}^{N} x_k e^{ik\omega} \right|^2. \tag{2.7}$$

This is also known as the periodogram of x, usually calculated using the Fast Fourier Transform.

The Drouiche statistic for estimating spectral variance is as follows,

$$W_N = \log \frac{1}{2\pi} \int_{-\pi}^{\pi} I_N(\omega) \mathrm{d}\omega - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log I_N(\omega) \mathrm{d}\omega - \gamma, \tag{2.8}$$

where γ is the Euler constant. If x is the output of a zero-mean Gaussian process then W_N is a normal random variable. When standardized, the pdf of W_N approaches the standard normal distribution,

$$\hat{W}_N = W_N \frac{\sqrt{N}}{\sqrt{\frac{\pi^2}{6} - 1}} \tag{2.9}$$

$$\hat{W}_N \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \tag{2.10}$$

2.3.2 Transforming the Ljung-Box statistic to the standard normal distribution

Later in this paper (Section 2.5.2), we compare between the Ljung-box Q statistic, which follows χ^2 distribution and the Drouiche $\hat{W}$ statistic, which follows the standard normal distribution. The Wilson-Hilferty transformation (Wilson and Hilferty, 1931) is a method for transforming a χ^2 random variable to a normal random variable, such that the transformed variable sits at the same percentile of the normal CDF as the original variable did with respect to the χ^2 CDF. Using the Wilson-Hilferty method, we transform the Q value obtained from Equation 2.5, to $\hat{Q}$, which follows the standard normal distribution,

$$\hat{Q}(x) = \frac{\frac{Q(x)}{m}^{1/3} - \mu}{\sqrt{\sigma}}, \tag{2.11}$$

where $\mu = 1 - (2/9m)$ and $\sigma = 2/9m$. This transformation enables us to compare the Ljung-box test results directly against the Drouiche test results, by comparing Q to $\hat{W}$, both of which are standard normal random variables when the input signal is GWN. We will report the JND value in Section 2.4 based on the standardized Q-value ($\hat{Q}$) in Equation 2.11 above.

2.4 LISTENING TEST

2.4.1 Psychometric evaluation method

Commonly practiced psychometric evaluation methods in the literature include the method of constant stimuli, the method of limits, and the method of adjustment (Guilford, 1954; Gescheider, 1997). These methods are used to determine perceptual thresholds (Guilford, 1954; Gescheider, 1997). In this paper we measure the spectral variance JND, which is the smallest spectral deviation from a reference state (GWN) that is audible to listeners.

In the method of constant stimuli (Guilford, 1954; Gescheider, 1997), each subject typically reports whether or not he or she notices a difference in each pair of stimuli presented. One sample in each pair is the reference stimulus and the other is the variable stimulus with adjustable intensity. This process is typically repeated hundreds of times with various amount of stimulus intensity in the variable stimulus in random order. The threshold is determined as the amount of intensity in the variable stimulus that is detected half the time. This method is known to be time consuming since listeners have to be exposed to all values of the variable stimuli. During our preliminary test, listeners complained of fatigue and was unable to complete the test using this method. Hence the method of constant stimuli was not a viable option.

The method of adjustment (Guilford, 1954; Gescheider, 1997) and method of limits (Guilford, 1954; Gescheider, 1997) are known to be more efficient since they allow us to find the threshold with smaller number of samples. In both methods, the subject is presented with two similar stimuli. The intensity in one of the stimuli is then gradually increased until the difference between the two is perceivable. This is called an ascending test. Alternatively, the subject may be presented with two stimuli that are very

different, and the intensity in one of the stimuli is slowly reduced until the difference is no longer perceivable. This is called a descending test. The difference between the method of adjustments and the method of limits is that in the method of adjustments the subjects control the intensity of the variable stimulus by themselves, while in the method of limits, the adjustment of the variable stimulus is automated or controlled by the experimenter.

There are two types of errors that often arise in the method of limits and the method of adjustment, called habituation errors and expectation errors (Guilford, 1954; Gescheider, 1997). Habituation error describes the problem where the subject becomes accustomed to giving one type of answer. For example, the subject makes a habit of answering yes when asked if he or she hears a difference between the two stimuli, resulting in underestimation of the threshold in descending trails and overestimation in ascending trials. Expectation error arises when the subject anticipates the threshold value, resulting in underestimation of the threshold in ascending trials and overestimation of threshold in descending trials. A common solution to eliminate these errors is by averaging the results from an equal number of ascending and descending trials. Hence for this listening test, we conducted two ascending and two descending trials. We then select lowest value from the two ascending trials and the lowest of the two descending trials and average those two numbers. In section 2.4.2, we further explain how our listening test procedure further mitigates both errors.

Additionally, we incorporated the staircase method (Cornsweet, 1962) in our listening test procedure. This modification allows the subjects to concentrate around questions that are in their threshold range and find their whiteness threshold efficiently. It allows subjects to complete all four sets of test before reaching fatigue. Due to this reason we chose to design the listening test to resemble the method of limits rather than the method of adjustment so that the staircase procedure can be automated.

2.4.2 Listening test procedure

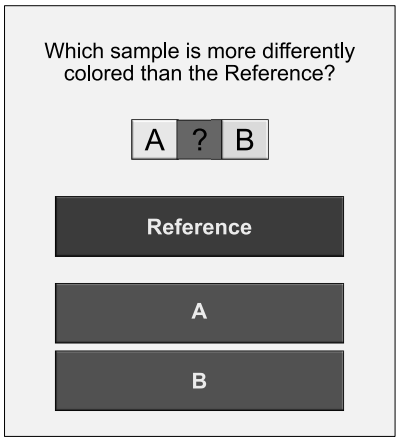

FIGURE 2.1: The user interface of iOS app used to administer listening tests. Subjects may tap and hold buttons A, B, and Reference in any order to listen to the audio files.

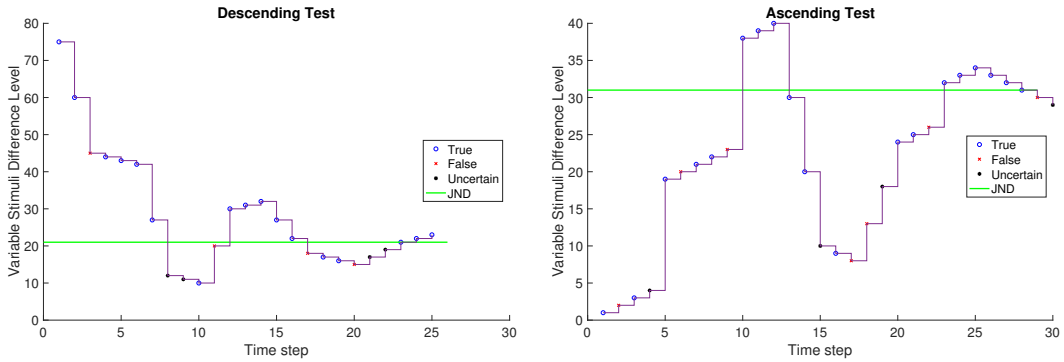

FIGURE 2.2: An example of a subject traversing the questions in descending (left) and ascending (right) staircase method of limits listening test. The y-axis represents the *difference* in the level of intensity between the variable stimuli: spectral variances α and β.

User interface

We created an iOS app to administer the listening tests. The user interface of the app is shown in Figure 2.1. The subjects were able to play three different audio files by tapping and holding the A, B, and Reference buttons. Each audio file is approximately 6 seconds long.

Variable stimuli

The Reference file is a Gaussian white noise denoted as $\mathcal{G}$. The A file is another GWN denoted as $\mathcal{K}$ with added variable spectral variance α, and the B file is the same GWN $\mathcal{K}$ with added variable spectral variance β. Hence, the variable stimuli in this listening test takes the form of spectral variance α and β. By definition, the Reference file $\mathcal{G}$ does not contain any stimuli or spectral variance. Both $\mathcal{G}$ and $\mathcal{K}$ are similar GWN sequences of the same length. We chose $\mathcal{G}$ and $\mathcal{K}$ such their Q values are similar (see Section 2.4.3 for details on how to generate them).

The condition for α and β in this test is that $\alpha = 0$ if $\beta > 0$ and $\beta = 0$ if $\alpha > 0$. Also, they can only take positive values since they represent the expected amount of squared deviation from a flat spectra. When either α or β is zero, then either file A or B is the unaltered GWN $\mathcal{K}$. Otherwise when either α or β is positive, then either file A or B becomes *colored* respectively. With this condition, only one of the files, A or B is colored at any given time. In Section 2.4.3 we further explain the procedure to generate these noise files.

Task

Subjects were tasked to indicate which one of the audio files, A or B, sounds more differently colored than the Reference file by choosing the answer at the top of the screen. In other words, to choose whether it is file A or B that is the colored noise (having positive spectral variance). Recall that only one of the files, A or B is colored at any given time. We did not limit their time to complete the listening test. Short breaks between trials were encouraged to delay the onset of fatigue.

Subjects were not required to listen to the entire 6s length and were free to repeat any audio files in any order. If the subject could not distinguish between A or B (i.e. perceive both has having the same color), he or she may indicate an 'uncertain' answer, denoted by the question mark '?' in Figure 2.1. This was done so that the candidates were not forced to give random answers.

Prior to the test, subjects were briefly instructed to focus on the color (texture) of the noise instead of plain, time-domain or temporal cues difference (differences in signal values over time) between the noise files. This was done because the concept of noise color is more subtle than temporal cues. In our informal test, we found that time domain differences are more easily perceptible. On the surface, two GWN signals may sound different despite having a flat spectra. This is because a GWN signal may carry certain temporal artifacts, such as abrupt changes in amplitude, short periods of time where the random variation in the spectrum temporarily accentuates a particular frequency to the point that we begin to have a sense of tiny elements of tonality. One could describe these various types of events as squeaks, clicks, or crackles throughout the noise signal. However the overall color or texture of the noise is white. Therefore, subjects who weren't familiar with the concept of noise colors were briefly trained to notice it. To avoid bias, we showed them samples of coloured noise signals that aren't used in the main listening test (other white, brown, blue, and pink noise samples). Once the subject is able to grasp concept of noise color, we allow them to begin the test. On average, most subjects only took a few minutes to understand and perceive noise color.

Also, due to this subtle notion of noise color, we added the Reference file as another example of GWN file instead of directly asking subjects to simply choose whether file A or B is the colored noise. The purpose of the Reference file was to guide and remind them of the perceptual impression of a GWN noise. They might have noticed that files A and B have different color, but without the Reference noise they may face difficulties in deciding which of the signals is less white (or equivalently, more colored). The prior brief training was done so that they may notice the notion of noise color in a given signal, but it wasn't enough to familiarize themselves with the absolute perception of whiteness.

Ascending and descending test procedure

Each subject has to complete two ascending and two descending tests. Subjects can either begin with descending or ascending test at random but will not do the same type of test consecutively.

For the descending test, the difference between α and β is set to be at maximum, which we call level 75. We assign at random which variance, α or β is nonzero. Each time the subject correctly identifies the file with higher spectral variance, we progressively reduce the level of difference between α and β and repeat the task. Otherwise, we reverse the direction and increase their level of difference. The test stops once the subject triggers the fourth direction reversal. Figure 2.2 (left) shows an example of how a subject progress through a descending test until a threshold is found. The speed of increase or decrease in stimuli levels is 15 levels at a time before the first reversal, 10 levels before the second reversal, 5 levels before the third reversal, and 1 level before the fourth reversal. This is indicated by varying step heights in Figure 2.2.

To ensure that reversal and the subject's JND is meaningful and not by chance, we allow reversal from higher level to lower level only when the subject correctly identifies the signal with more variance in at least three consecutive levels, and reversal from lower level to higher level when subject fails to do so. In Figure 2.2 (left) we show one possible scenario where reversal from lower to higher level is canceled because the subject was able to give the correct answer in three consecutive levels although he gave the wrong answer at level 45.

Similarly in ascending test, the difference between α and β is set to be at minimum, which we call level 1. Figure 2.2 (right) shows an example of how a subject progress through a ascending test until the fourth reversal is triggered.

By randomizing the assignment non zero variance and by requiring the subjects to correctly identify the button with colored noise for at least 3 consecutive levels before reversal, we reduce the effect of errors due to habituation and expectation. The listening test procedure mitigates habituation errors because the correct answer sequence is random, therefore the subjects could not have identified the colored noise correctly if the difference was not noticeable. It is also impossible for expectation error to artificially lower the JND we obtain from this test because the listeners cannot report the threshold below level of variance where they can correctly identify the colored noise. However, it is possible for expectation error to artificially raise the JND. For this reason, we selected the lowest ascending and lowest descending threshold values from each listener and set his or her JND as the average.

2.4.3 Generation of audio files

Generation of GWN $\mathcal{G}$ and $\mathcal{K}$

Both $\mathcal{G}$ and $\mathcal{K}$ are 2^{18} samples long GWN (5.94s at 44100 Hz sample rate). The Q-value of GWN follows the χ^2 distribution (Ljung and Box, 1978), meaning that sequences at the 50th percentile of the statistic have $Q = m$, where m is the maximum autocorrelation, in samples. Therefore to generate each noise files $\mathcal{G}$ and $\mathcal{K}$, we produced many 2^{18} samples long random sequences derived from standard Normal distribution until we found one such that $\|Q - m\| < \epsilon$, where ϵ is the mean absolute difference of the χ^2 distribution with m degrees of freedom. This is important because the spectrum of Gaussian noise is random and some observations fall far from the mean value of spectral variance. We want the both $\mathcal{G}$ and $\mathcal{K}$ to have Q near m to ensure that the random sequence we select is a typical example of GWN and not an outlying case. Hence, both $\mathcal{G}$ and $\mathcal{K}$ are GWN but they do not have identical individual sample values.

Generation of colored noise signals

The colored signals for this listening test is GWN $\mathcal{K}$ that is processed through a randomised finite impulse response (FIR) filter with adjustable spectral variance α to produce audio files A and β to produce audio files B. To create these filters, we generated the Fourier series of a signal whose power spectrum is randomized with the desired mean and variance and use the inverse Fourier transform to produce the filter kernel. The length of the kernel is 2^{14} samples. We convolved $\mathcal{K}$ with the kernel to produce

A and B files. We then compute the Ljung-box statistic of the colored noise signals to ensure that the randomization of the filter produced a result with the desired $\hat{Q}$ value.

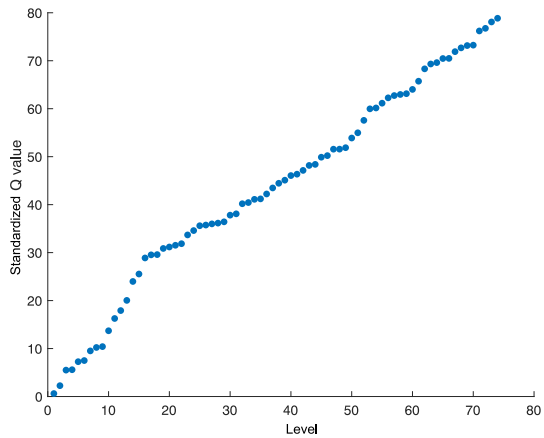

FIGURE 2.3: The plot of corresponding average standardized Q-values of the colored signals with stimuli level 1 to 75 used for the 4 trials of 2 ascending and 2 descending trials.

When not taking the value of zero, spectral variance α and β can be varied in the range of 0.5 to 15.5 dB. We precomputed these signals with various spectral variance in 0.15 dB increment between 0.5dB and 15.5dB so that they can be accessed directly at runtime to speed up the test. In total we have 75 different levels of spectral variance. The first level corresponds to 0.5dB of variance, the second level corresponds to 0.65 dB of variance and so forth. Finally, the 75th level corresponds to 15.5dB of variance.

It is also important to note that we precomputed four different sets of colored signals, each set containing 75 signals from 75 different levels of added spectral variance. Similar to the generation of $\mathcal{K}$ and $\mathcal{G}$, we ensured that the difference in $\hat{Q}$ values of each colored signal with the same spectral variance level across four sets is smaller than ϵ. This is to ensure that the individual sample values of signals in the same level in each of the four tests (two ascending and descending) are different despite having similar $\hat{Q}$ amount. Figure 2.3 shows the average $\hat{Q}$ values that corresponds to these 75 different levels of added spectral variance in all 4 trials. Higher $\hat{Q}$ means that the signal is less white or more colored. For example, when α is set to be at level 10 (2dB spectral variance) in the first trial, the app associates button A with the precomputed signal at level 10 in the first precomputed set of questions designated for the first trial (which its $\hat{Q}$ value is 13.2) and button B with GWN $\mathcal{K}$.

Frequency range of all test signals

The GWN files $\mathcal{G}$ and $\mathcal{K}$ generated have spectral energy between 0 and 22050 Hz. However, the FIR filters only vary the frequencies in the range from 50 Hz to 16500 Hz. We set the lower limit at 50Hz in case of hardware limitations in the ability of producing frequencies below 50 Hz as accurately as higher frequencies. The upper limit is based on the assumption that human hearing deteriorates with age and that older listeners may only hear up to 14000Hz - 16000 Hz. Filtering out higher frequencies helps to

ensure that age differences do not affect the test results (Stelmachowicz et al., 1989; Crocker, 2007).

To compute the $\hat{Q}$-values of the noise files, we set the maximum autocorrelation lag to $m = 44100/50 = 882$ samples. This is so that the its value will not be influenced by frequencies below 50 Hz, since the relation between lag time and its corresponding frequency is as stated in Equation 2.12.

2.4.4 Equipments

We installed and ran the app on an iPad Mini. We used (ER 400 SR) studio reference in-ear headphones and an amplifier. To ensure that no unintended filtering from the equipment (headphones and iPad) affected the test, we measured the frequency response output of the headphones by playing a GWN and recording the output from the headphones using a reference microphone (G.R.A.S. 46 BD 1/4" CCP Pressure Standard Microphone Set), a conditioning amplifier (Bruel & Kjaer Nexus), and an audio interface (MOTU UltraLite MK4). We then applied the same inverse filter to all of the recordings to flatten the frequency response. The tests were conducted in a very silent environment, a carpeted small indoor meeting room (approximately 2m by 3m by 3.5m) located inside a quiet office The doors were closed at all times and the air conditioning was turned off during the test. Residual ambient noise was reported as inaudible once the in-ear headphones were plugged into the ear canals.

2.4.5 Subjects

A total of 50 subjects, 24 females and 26 males between 20 and 40 years of age participated in this listening test and received remuneration. Similar-sized groups of test subjects have been used to establish JND in psychoacoustics related fields, for example, in (Martellotta, 2010) and (Buck et al., 2012). None of the test subjects reported any hearing impairments. Nine of them have experience with audio recording and mixing and 28 out of 50 are musicians. One subject decided not to complete the test when he found that he was unable to grasp the idea of noise color during training. Therefore we present results from the remaining 49 subjects in the next sections.

2.4.6 Test results

Statistical results

The subjects took on average 30 minutes to complete all four trials of the test. Figure 2.4 presents the answers from each of the 49 subjects, two results (one from the ascending and another one from the descending test that have the lowest JND value) per subject. In total, there are 108 results displayed in Figure 2.4. The y-axis indicates the degree of whiteness, presented in terms of $\hat{Q}$ values (standardized Q-values, computed from Equation 2.11 from a given Q) of the colored signal (refer to Figure 2.3), that each subject encountered during traversal via the staircase method. The circle, cross, and star labels indicate true, false, and uncertain answers respectively. We can immediately notice in Figure 2.4 that the answers above $\hat{Q} = 50$ and below $\hat{Q} = 20$ are sparser. Most of the answers are concentrated around levels with $\hat{Q}$ between $20 - 50$. The staircase method leads to this phenomenon, as it allows the candidates to quickly skim over levels that

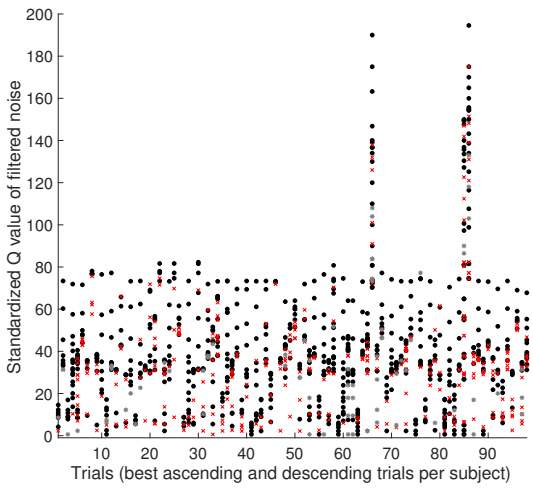

FIGURE 2.4: The result of best ascending and best descending trials for all 49 subjects plotted against standardized Q-value as computed in equation 2.11. Filled circles indicate true answers, x indicates false answers, and * indicates uncertain answers.

are not near their JND, which is levels with $\hat{Q} > 50$ and $\hat{Q} < 20$. Outliers are present in trial 66 made by subject 33 (descending trial), and in trials 85 and 86 (both ascending and descending) made by subject 43.

Almost all participants correctly identified the colored noise with high $\hat{Q}$-values, $(\hat{Q} > 56.5)$, that is when the colored noise was far less white than the Reference file. On the other hand, there is a relatively even mixture of true, false, and uncertain answers for $\hat{Q}$ between 0.4 and 15. This indicates randomness in their answers, and it shows that the subjects are not able to clearly distinguish the colored noise from the GWN in these low levels of spectral variance. The staircase method leads test subjects to listen to more levels with $\hat{Q}$ values between 27.8 - 43.3, indicated by denser points around this range in Figure 2.4. This is because for most subjects, the JND lies in that range and the staircase method requires them to pass over the JND value three times, reversing direction thrice.

Figure 2.5 shows the histogram of the spectral variance threshold values of all 49 subjects. Recall that each subject was required to do 4 trials: 2 ascending and 2 descending trials. As mentioned in the previous section, the JND for each test subject is found by averaging the single-trial JND from lowest descending and the lowest ascending trials from each subject. The median JND from 49 subjects is $\hat{Q} = 33.8$. The mean, minimum, maximum, and standard deviation is shown in Table 2.1.

The large standard deviation is due to the presence of two outliers shown in figure 2.5, with JND of $\hat{Q} = 101.8$ and $\hat{Q} = 151.9$ respectively. These two subjects were unable to correctly identify the colored signal even at level 75 spectral variance ($\hat{Q} = 83.2$), where the difference between audio file A and B is at its maximum. These $\hat{Q}$ values correspond to a spectral variance of 22 dB and 30 dB, respectively. The remaining 47 subjects did not have any difficulty finding their JND threshold below 15dB of spectral variance. The responses from the two outlying test subjects are shown in Figure 2.4, trials number 66 (subject 33), 85 and 86 (subject 43). The first boxplot on the left in

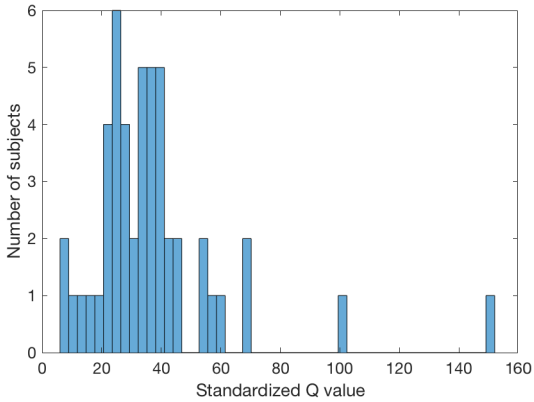

FIGURE 2.5: Histogram of spectral variance threshold values of all 49 subjects in terms of $\hat{Q}$ value from best ascending and best descending per subject. The standard deviation for this histogram is 23.6

Figure 2.6 shows more clearly the presence of these outliers. We presented them with additional colored signals with much higher $\hat{Q}$ values above 87 until we found their JND threshold. On the opposite end of the spectrum, two subjects who claimed to be have perfect pitch scored very low thresholds at near $\hat{Q} = 4.82$ which corresponds to a spectral variance of 1.2dB.

Based on our informal observations of the test subjects, we believe that the ability to hear a difference for lower difference levels is not only a function of hearing ability. With the tremendously wide range of JND results coming from a group of people who all have normal hearing ability, it seems more likely that the variance also has a psychological explanation. We observed an obvious correlation between the JND score and the attitude with which the listeners approached the test. The majority of the subjects showed considerable effort in completing the test, but some listeners expressed dislike for the sound of the noise after a few trials, and had unusually high JND at the last trial. We attempted to reduce the effect from this problem by selecting best ascending and best descending trials as their JND. A couple of listeners became involved in an ego-contest to show off their listening skills, and tried to do the test to the best of their ability. Not surprisingly, they made up part of the candidates who were able to consistently and correctly identify the colored noise at the lower levels of spectral variance and also took the longest time to finish the test. They expressed that they felt competitively motivated to achieve a better JND score. On the other end of the spectrum, a few subjects said that they did not feel this test was meaningful or significant. These type of candidates tend to give random answers at lower levels and only correctly identify the colored noise when the difference level is extremely high, such as the case with subject 33 and subject 43. They form the outliers. For this reason, we believe that we will overestimate the JND if we set it to be the median JND value from all 49 subjects. We feel that the lower values in the results are the more relevant indication of the actual JND for spectral variance, and therefore the overall JND will be more accurate if we eliminate the outliers from the dataset.

The simplest methods for removing outlying scores work by trimming both extreme

Type	N	Median	Mean	Min	Max	σ
All	49	33.84	37.29	6.43	151.88	23.6
MAD (3)	42	30.35	30.10	6.43	53.71	10.49
MAD (2)	38	31.79	31.23	13.66	46.28	8.16
MAD (1.48)	33	32.63	31.66	31.48	43.74	6.53
10% Trimmed	41	33.84	34.12	16.60	59.67	10.43
10% Wind	49	33.84	34.42	13.66	59.67	13.13

TABLE 2.1: Median, mean, min, max, and standard deviation of the personal JND for all 49 subjects and of the remaining subjects after applying various methods to eliminate outliers. N is the number of data points left after the outlier trimming methods have been applied.

values in the dataset, or by Winsorizing (Searls, 1966) these extreme values. The difference between trimming and Winsorizing datasets is that instead of simply discarding the top and bottom p percentile, in Winsorization method we replace the respective extreme data with values from the pth top and bottom percentiles. A more robust way to exclude outliers is by using the median absolute deviation (MAD) (Hampel, 1974). This method can be used even for datasets that are not known to be normally distributed. As a cutoff, we can typically use a consistency scale factor of 2.0 to 3.0. If we assume that the threshold data is normally distributed then a consistency scale factor of 1.4826 can be used. Table 2.1 lists the median, mean, min, max, and standard deviation values of the threshold dataset after applying various methods to eliminate outliers, along with the number of data points left. The boxplots in Figure 2.6 graphically show the distribution of the data after various trimming methods. We can see that the variance is tremendously reduced when the outliers are eliminated, especially when MAD method is used.

Spectral variance JND

By definition, the median value can be translated to be the JND of spectral variance since at this Q-value, the colored noise is correctly distinguishable from the GWN by at least half the subjects (Fechner, 1966). From Table 2.1, the median value ranges from about $\hat{Q} = 30.35$ to $\hat{Q} = 33.84$, depending on the method used to eliminate outliers. This corresponds to spectral variance stimuli between 6.8dB to 7.4dB.

Feedback and analysis

We conducted several informal experiments prior to the test in section 2.4 and concluded that 4 trials is the maximum amount the listeners can comfortably finish before reaching fatigue. During this preliminary round, our test subjects expressed some fatigue after 20 minutes (completed 4 trials without breaks), and more visible fatigue and discomfort after 30 minutes (completed 6 trials without breaks). We then encouraged them to take breaks in between trials and even then, most subjects show some signs of struggle in completing the 5th trial. We believe that our ability to collect meaningful

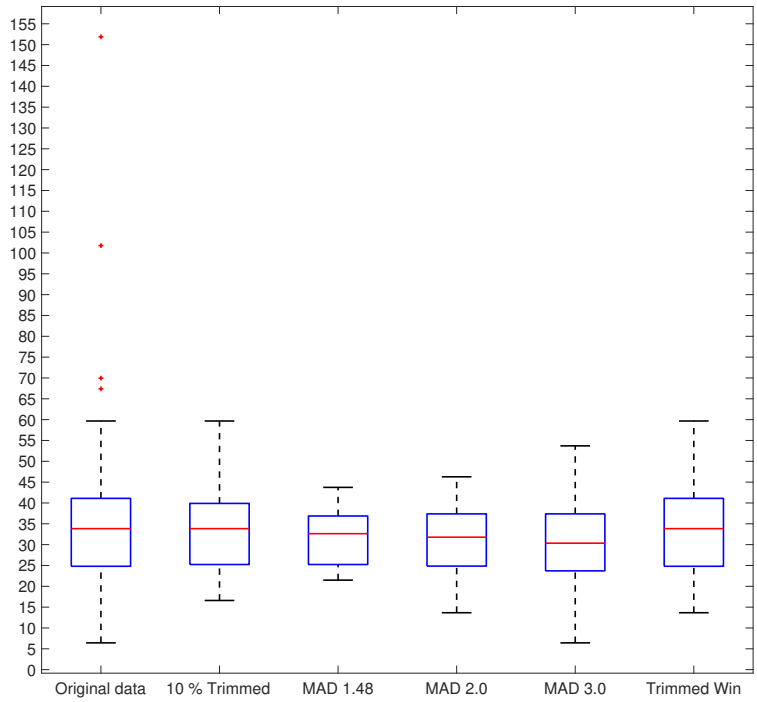

FIGURE 2.6: The boxplots of listening test results from candidates after various trimming methods to eliminate outliers. The y-axis corresponds to the $\hat{Q}$ value of the colored noise presented to the subjects. The original result from all 49 subjects that still contains the outliers is shown at the leftmost plot.

results diminishes when the listening test subject is tired because under fatigue conditions he or she becomes more easily confused and unable to correctly distinguish the colored noise from the GWN. For the listening test in section 2.4, we encourage the subjects to take their time in completing the test. They typically take very short breaks in between trials, thus leading to an average of 30 minutes completion time for 4 trials.

It is impossible for the subjects to artificially lower their JND since they were tasked to correctly identify the randomly assigned colored signal on button A or B three times in a row before we record their JND. However, this does not prevent the listener from reporting the threshold higher than the real JND. Some listeners appear to stop making an effort to hear the difference between the two signals A and B when they reach the level of variance where they think the JND ought to be located based on experience from a previous trial.

At first, it seems logical to present the subjects with files A and B and asked them to indicate whether there was a difference. We thought that such test would have been much shorter and the subject could have completed more trials before reaching fatigue, thus reducing the effect of habituation error that may surface. However due to the subtle perception of noise color (especially noise whiteness), the differences in temporal cues were overwhelming in comparison. As discussed earlier in section 2.4.2, even two GWN signals could certainly sound different despite having the same color. The subjects will almost certainly indicate that there is always a difference between the two signals and it defeats the purpose of the listening test. The test was then modified to require the subject to correctly identify which of the two files were colored and that the correct answer is randomized. However this modification was not good enough because subjects who are not experienced with the notion of noise color do not know which of the file is less white, despite fully realising that they are differently colored. In the preliminary round, we found that the subjects actually took longer time in completing the test due to this confusion. The presence of the Reference file served as a useful guidance on how the color of an ideal GWN is perceived, and in fact causing the duration of the test to be shorter. Most subjects constantly compare A and B to the Reference file and quickly select the file that they feel is perceptually 'further' away from the Reference file. A few of more experienced subjects, such as those who are experts in audio mixing did not make use of the Reference file and were able to identify the coloured noise because they were already familiar with the notion of white noise.

We did not find a strong correlation between musical background and the spectral variance JND. Subjects with musical background do not necessarily have a lower threshold or vice versa. In fact, the one subject who failed to complete the test is experienced in choir yet another subject that scored one of the lowest JND threshold is an experienced pianist. However we found that the subjects who are experienced with audio mixing and recording have their threshold slightly below the median JND. Further study needs to be done with more related candidates to establish if correlation exists between their experience as an audio engineer and their spectral variance thresholds.

We also would like to add a note that not all listeners who had higher JND approached the test negatively. It appeared that they truly had higher JND values than the average despite trying their best and having what it seems to be a normal hearing ability. Conversely, not all the listeners who scored very low JNDs were competitive (showing excessive effort) or took a long time to complete the test. There were a couple

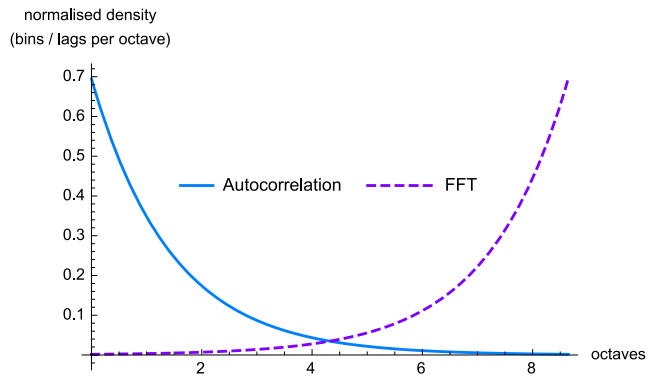

FIGURE 2.7: The plot of the density of lag times and FFT bins per octave interval on the log scale frequency spectrum.

of listeners who seemed to be able to distinguish correctly the colored noise at the lowest levels effortlessly. Investigating the correlation between hearing ability, attitude, and spectral variance JND is beyond the scope of this paper, but it points to a possible direction for further study.

2.5 UNEVEN SPECTRAL EMPHASIS

2.5.1 Disagreement between Q and W

Definition 2.5.1. Disagreement - for two sequences x and y and two spectral variance measures W and Q, if $W(x) > W(y)$ and $Q(x) < Q(y)$, then Q and W *disagree* about which of the sequences has the greater spectral variance.

The Ljung-Box test, although strongly correlated, frequently disagrees with the Drouiche statistic on pairs of signals with similar but high amount of spectral variance. This occurs because the underlying time-frequency transformations give different emphasis to different parts of the spectrum. The computation of Q in Equation 2.5 requires autocorrelation that depends on lag time k. In autocorrelation, the frequency corresponding to lag time k, denoted by f_k is as follows,

$$f_k = \frac{f_s}{k+1}, \tag{2.12}$$

where f_s is the sampling rate. By distributing the autocorrelation measurements r_k in Equation 2.5 unevenly over the spectrum, emphasis is effectively given to variance at lower frequencies and the high frequencies are de-emphasized. On the other hand, the computation of W in Equation 2.8 utilizes Discrete Fourier Transform (DFT). The bins of DFT are evenly spaced over the linear frequency spectrum between Direct Current (DC) and Nyquist. Figure 2.7 shows the density of lag times k and FFT bins per octave in the log-scale frequency. We can observe that the Ljung-Box test has excessive emphasis on the lower parts of the spectrum and the Drouiche test excessive emphasis on the higher parts of the log-scale spectrum.

At this point we consider the question, what distribution over the frequency spectrum, of lag times or FFT bins, would be optimally white from a perceptual standpoint? It is well known that the human hearing apparatus is not linear, and therefore the linear frequency scale is not relevant to the perceptual sense (Davis and Jones, 1989). The scales that may be more relevant to the auditory sense are the Bark scale (Zwicker, 1961) or the logarithmic scale.

In the next section we describe a listening test result that indicates that neither the Ljung-box statistic nor the Drouiche statistic is more strongly correlated with human perception.

2.5.2 Evaluation by listening test

Autocorrelation is used to compute Ljung-Box Q-value, while the FFT is used to compute the Drouiche W-value. As explained in the previous section, both methods place uneven emphasis across the audible spectrum. Since we set $m = 882$ to account for autocorrelation above 50Hz, more than half of these lags used to compute Q correspond to frequencies between 50Hz and 100Hz. The density of lag times per Hz decreases as f increases (see Figure 2.7). In contrast, the frequency f' that corresponds to bin number i of FFT of a signal with length N is $f' = i(F_s/n)$. Although this gives equal emphasis in linear frequency, it is uneven with respect to the more perceptually relevant log scale and bark scale spectra. An increase in spectral variance in the frequency range from 50 to 100Hz will affect the Q-value of the signal more than the W-value.

As the spectral variance α or β increases, we can see an increase in cases of disagreement between $\hat{Q}$ and W values. Figure 2.8 shows strong correlation between W and $\hat{Q}$ but a degree of disagreement is also visible at high values of $\hat{Q}$ and W.

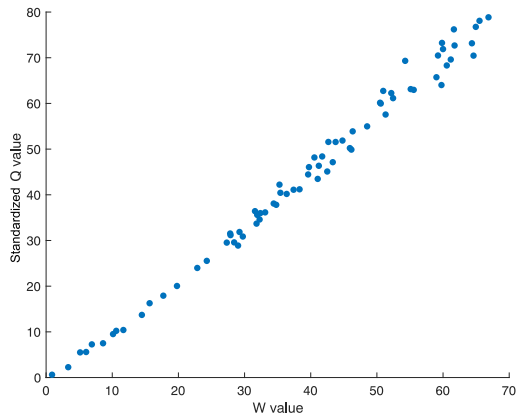

FIGURE 2.8: Standardized Q values and W values of all colored signals used in one of the four trials in the JND listening test. Some degree of disagreement can be observed around the top-right region of the plot.

Therefore, we conducted a second listening test with a smaller pool of test subjects to determine which of the metrics, Q or W, is more correlated with human perception. A total of 10 subjects participated in the second test, 4 females and 6 males with ages ranging from 20 to 40 years. These subjects also participated in the JND listening test

explained in Section 2.4. In this second test, subjects were presented with 30 pairs of signals. For each pair in this test, we selected signal A and B such that the Q and W values disagree about which of the two signals has greater spectral variance. The W and Q values for each pair are shown in Figure 2.9. Each subject was asked to choose which signal, A or B, is more colored (less white). We also present them with the white Gaussian noise signal $\mathcal{G}$ used in the previous listening test as a reference point at all times. No time limit is imposed and subjects were free to replay any signals in any order. The purpose of this test is to determine if the listener's answers are more correlated with W or with Q in cases where W and Q disagree with each other.

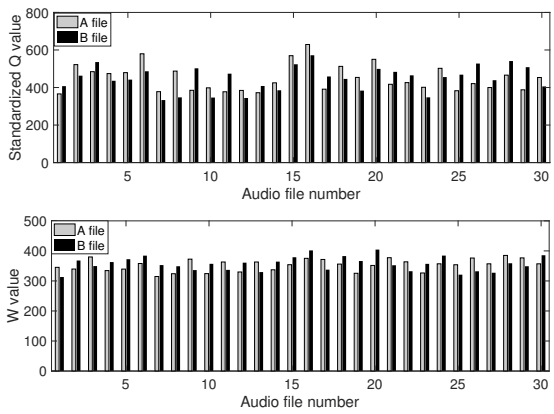

FIGURE 2.9: The top graph shows $\hat{Q}$ values (standardized Q-values) for thirty pairs of audio signals. The bottom graph shows standardized W values for the same thirty pairs of signals. Note that in each pair the W and $\hat{Q}$ graphs disagree about whether the A or the B signal has greater spectral variance.

After each subject is done, we counted the number of pairs of signals for which the listener agreed with Q, meaning that the subject selected the signal with higher Q value as more colored. We also counted the number for which he or she agreed with W. Figure 2.10 shows the results of the second listening test. From the figure, we can see that there is no consistent agreement between the their answers and either W or Q.

The graph in Figure 2.7 is a possible explanation of that result. In Figure 2.7, we can see that the normalised weights given to each octave of the frequency spectrum by autocorrelation and FFT-based methods are exact mirror images of each other in the log spectrum. None of the methods has linear weights in the logarithmic scale, that is known to be more related with perceptual sense (Davis and Jones, 1989)

2.6 SUMMARY

There are many applications where it is desirable to achieve a signal, residual, or output that is as close as possible to ideal GWN. In the context of building a digital reverberation network, for example, we typically begin by designing a lossless prototype and tuning its parameters to achieve a white-noise-like impulse response. In the process of optimising those parameters, it would be helpful to know how close to GWN we have

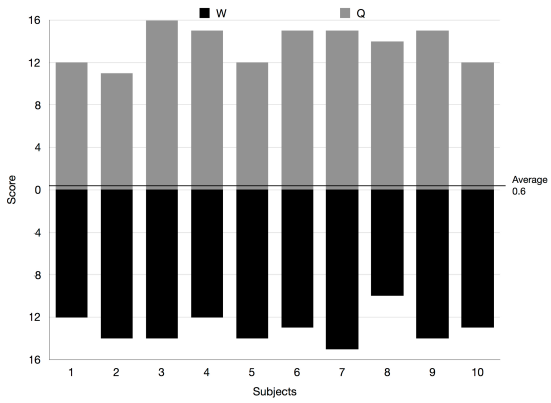

FIGURE 2.10: Results of second listening test, checking for agreement between the listener's perception and the spectral variance measures W and Q. The bars and the vertical axis indicate the number of signal pairs for which each listener agreed with W (bottom) or Q (top). The numbers on the horizontal axis indicate the ten listening test participants.

to get before the difference becomes inaudible. In this work we studied the perceptual threshold, or also known as the JND value for spectral variance to find out how far from ideal GWN a signal can be before the coloration becomes audible. Results from all 49 candidates who participated in the test show that the JND value for spectral variance is $\hat{Q} = 30.35$. As shown in table 2.1, this value can vary slightly between $\hat{Q} = 30.35$ and $\hat{Q} = 33.84$, depending on the methods used to trim the outliers. Since we believe the results may be slightly biased toward the high end due to lack of enthusiasm on the part of some test subjects, we consider the lower bound estimate to be the better estimate. We say this because if in some signal processing application we knew that a noise we intend to sound white would in fact sound coloured to a significant fraction of listeners, we would consider that noise as not perceptually equivalent to GWN. In the next Chapter, we use this JND value to evaluate whether ray-tracing delay lines cause any noticeable coloration to a lossless FDN.

In Section 2.5 we show that Ljung-Box method places uneven emphasis on particular parts of the frequency spectrum, mainly concentrating on the lower frequencies. However, Drouiche Test and SFM place emphasis on the upper part of the spectrum much more than the human hearing apparatus does. Since these methods do not place the same amount of emphasis on particular parts of the frequency spectrum, they may have contradicting results that become more apparent as spectral variance increases. We found no indication that neither the Ljung-Box statistic or the Drouiche statistic is significantly more correlated to human perception. As future work, we suggest further investigation to this matter by developing a spectral variance measure that places even emphasis across a perceptually relevant frequency scale such as the Mel scale (Stevens and Volkmann, 1937) or the Bark scale (Zwicker, 1961).

2.7 FUTURE WORK

In the literature we sometimes see the JND used as a unit of measurement, as if a value of three JND could be understood to be just audibly different from a value of two JND, as three JNDs is one JND higher than two JNDs (Hak, Wenmaekers, and Luxemburg, 2012; Wendt, Par, and Ewert, 2014a). However, in the case of spectral variance we are not convinced that this would be accurate. For a future study we recommend conducting listening tests to locate the $\hat{Q}$ and $\hat{W}$ values for 1 JND above 1 JND, which means the minimum audible increase in $\hat{Q}$ above 30.35 for which listeners hear a difference. Similarly for 1 JND above 2 JND and so on, so that we can estimate a curve that expresses spectral variance in a unit that is linear in a perceptual scale. This would be helpful because other measures of spectral variance exist that are not linear with respect to Q and W and it is not clear which of them is a better indication of human perception.

Chapter 3

Whiteness of Lossless FDN Output

3.1 ABSTRACT

Ideally, when an impulse is fed to a lossless FDN, the output should resemble Gaussian White Noise. This means on average there are equal amount of frequency component across the spectrum. The output of a lossless FDN is heavily dependent on its delay length (Jot and Chaigne, 1991). A wrong combination of delay lengths in the FDN may heavily colorise its output, and hence resulting in an unpleasant reverberation effect. In Chapter 1, we established the physical significance of signals in FDN as the basis for our real-time binaural reverberation algorithm. In this Chapter we present the result on whether by setting the FDN delay lengths into the lengths of first order reflection paths (we call this *ray-tracing delay line*) will affect the whiteness of its output, if the FDN is set to be lossless. Recall that in Chapter 2, we established the JND of perceptual whiteness, which is the amount of colouration that can be added until a Gaussian White Noise is no longer perceived as being white. We compared the SFM value of the lossless FDN with this JND and we found that the lossless output of the FDN after we set the delay lengths into the paths of first order reflections is still perceivably white, which is below the spectral variance JND value.

3.2 BACKGROUND

There are infinitely many ways to set FDN delay lengths, but there are certain ground rules. In this section we list three most important properties of FDN delay lines, especially when they are used in room modeling systems to model the late reverberation tail of RIRs.

3.2.1 Relationship between delay line lengths

When Schroeder first introduced the idea of artificial reverberator using delay lines (Schroeder, 1962), it was mentioned that delay lines that are mutually prime are preferred in the design of the reverberator structure. A set of mutually prime numbers do not have common divisor that are larger than 1, e.g: $\{10, 13, 19, 27\}$. The reason why this is important is that we would like to spread out the output of the reverberator as much as possible so that the system may produce signals that sound random and natural. We do not want to have a structured, predictable output.

We can see why using delay lengths that are not mutually prime is a problem. Consider an FDN of size 3, with delay lengths of $\{4, 8, 12\}$, and that we input an impulse to all of them at the same time. The first sample is produced 4 unit times later, from the shortest delay line with length 4. Then, this output is fed back to all of the three delay lines in some proportion depending on the mixing matrix. We observe no output at the fifth, sixth, and seventh unit time. At the eighth unit time, we have output from both delay lines with length 4 and 8. Similarly, we observe no output at the ninth, tenth, and eleventh unit time. At the twelve-th unit time we have output from all three delay lines, followed by no output at the thirteen-th, fourteen-th, and fifteen-th unit time. In the end what we have is a completely predictable, staggered outputs per four unit time (since their common divisor is four), and not outputs produced with seemingly random timings from the reverberation. It repeats itself at their common divisor value. If the delay lengths are mutually prime, then we maximize the number of samples that the FDN produces before they repeat. They will eventually repeat at the unit time equal to their least common multiple, which is the multiplication of all the delay lines length if they are mutually prime, e.g: delay lines of lengths $\{10, 13, 19, 27\}$ will repeat at the $(10 \times 13 \times 19 \times 27)^{th}$ unit time.

3.2.2 Setting the lengths of individual delay lines

When FDN is used to model a room, the average delay line lengths should correspond to the mean free path $\bar{d}$ of the room. The mean free path is the average distance that sound waves can travel freely through the air before encountering physical obstacles such as walls and objects in the room. This quantity is approximated by Sabine as,

$$\bar{d} = 4\frac{V}{S}, \tag{3.1}$$

where V is the volume of the room and S is the total surface area of the room (Smith, 2010).

3.2.3 Setting the total length of delay lines and FDN size

We also need to ensure that the size of the FDN and the sum of the delay line lengths are high enough so as to achieve a high mode and echo density. High echo density means that the number of echoes per second are high enough to mimic that of the dense late reverberation tail of a room impulse response. In (Jot, 1997), Jot stated that FDN sizes between 8 to 16 are sufficient to create natural reverberation. At that time, computational power was limited. Current technology no longer has to be confined to FDN sizes below 16 since it is able to run even larger size FDNs (32, 64, and even 128) in real time and create even higher (smoother) echo density. Note that depending on the room modeled, a higher echo density is not always desired. For empty rooms, smaller rooms have high echo density while larger rooms have lower echo density.

Mode density is the frequency domain counterpart of echo density. Correspondingly, it is suggested by Schroeder (Schroeder and Logan, 1961) that a mode density of 0.15 per Hz is sufficient to produce natural sounding reverberation. The echo density is proportional to t^2, where t is time, hence after some time humans are unable to perceptually distinguish individual reflections and the echoes can be approximated as stochastic process (Smith, 2010). Similarly, the mode density is proportional to f^2 where f is frequency, and so at higher frequencies the modes appear random.

If we do not model high enough mode density then some frequencies will stand out among the rest and result in unnatural ringing sound. A suggestion for the amount of sufficient mode density is,

$$M = \sum_{n=1}^{N} M_i, \tag{3.2}$$

$$M > 0.15 \times RT_{60} \times f_s, \tag{3.3}$$

where M_i is length of delay line i, N is the total number of delay lines in the FDN, RT_{60} is the reverberation time in seconds, and f_s is the sampling rate. The complete reasoning behind the equation above can be found in citeSmith2010Physical. Intuitively, the FDN is a type of feedback control system with order [1] M, and therefore having M poles on the unit circle for FDNs without attenuations[2] (lossless). Modes represent the behavior of the system due to the poles, which can be directly observed/heard in time-domain when we play the signal or in the frequency domain when we take the Fourier Transform of the signal. There is one mode per pole, hence for a signal with sampling rate of f_s and a uniform mode distribution among all frequencies, its mode density is M/f_s per Hz.

3.2.4 Motivation

All these requirements for delay line lengths, delay line sum, and relationships between delay lines are imposed such that the FDN is colorless (Schroeder, 1960). This can be

[1] The order of a system is the amount of individual elements that can be stored in the system at any given time.

[2] FDN attenuation coefficients shift the poles to be inside the unit circle and therefore cause the FDN to loose energy.

achieved with even distribution of modes in the frequency domain hence the frequency response appears random with no emphasis on particular frequency region. In other words, an impulse input to a well-designed lossless FDN should produce an output that resembles natural (Gaussian) 'white' noise. Recall in Chapter 4 that 'white' noise refers to a signal which *expected* frequency response is a constant (or flat).

In (Smith, 2010), it was stated that there is no need to mathematically model too many echoes per sample or too many resonance than what human perception can comprehend. This means that there is a certain limit of perceptual 'whiteness' (we call this spectral variance JND in chapter 4), such that there is no need to spend more time in coming up with the best delay lines design for colorless FDN if the 'whiteness' of the FDN is already below that spectral variance JND that we found in Chapter 4.

There are a lot of suggestions on how to set the delay lines of an FDN so that it is perceptually colorless, but they are mainly circulating in the Internet due to suggestions or personal preference from various people. Some methods are way more complicated than the other and it was not clear whether such degree of complexity is (perceptually) necessary. There are more constraints on setting these delay lines when the FDN is used to model the late reverberation tail of an RIR, since it has to conform to the mean free path of the room modeled. We have yet to find a single literature that collate and systematically evaluate these methods. We use the JND we found in Chapter 4 as a benchmark.

3.3 PRIOR WORK

3.3.1 Spectral Variance JND

One of the prior works for this chapter is the establishment of spectral variance JND explained in Chapter 2. This JND value will be used as a benchmark when evaluating various methods for FDN delay line settings. The reason for using this JND as a benchmark is that we can justify more clearly whether there is a need to spend more computational power to come up with a more colorless FDN if it is doubtful whether such improvement is going to be noticeable. The lower bound JND found in Chapter 2 for spectral flatness variance is $\hat{Q} = 30.35$.

The conversion from $\hat{Q}$ to SFM is,

$$\hat{\Xi} = e^{-(\hat{Q} * \frac{\sqrt{\pi^2/6-1}}{\sqrt{2048}} + \gamma)}, \tag{3.4}$$

where e is Euler's number, γ is the Euler–Mascheroni constant, and therefore $\hat{\Xi} = 0.3288$.

3.3.2 FDN State-Space Representation

There are two ways to obtain the frequency response of an FDN, and subsequently check whether it is colorless. Firstly, is by plotting the transfer function of the FDN obtained from its state-space equation.

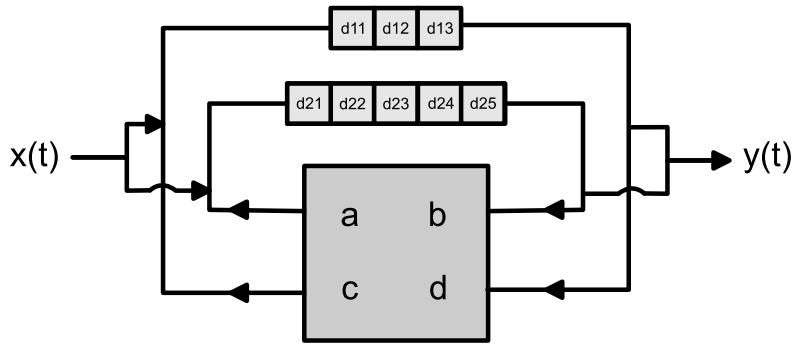

FIGURE 3.1: An example of an FDN with two delay lines of lengths 3 and 5 respectively.

The state-space representation of a system with single input and single output is given by two equations,

$$\dot{\boldsymbol{u}}(t) = \boldsymbol{A}\boldsymbol{u}(t) + \boldsymbol{B}x(t) \tag{3.5}$$

$$y(t) = \boldsymbol{C}^T\boldsymbol{u}(t) + D, \tag{3.6}$$

where $\boldsymbol{u}$ is a vector representing the states in the system, $\dot{\boldsymbol{u}}$ are vectors of the system state in the next time step, A is a square state matrix, B is the input vector, C is the output vector, and D is the direct transition constant, and $y(t)$ and $x(t)$ are scalar output and input respectively. A, B, C, and D are all constants.

We can represent a single delay line of length n (without any attenuation) in state-space form as the following two equations,

$$\begin{bmatrix} \dot{d}_1 \\ \vdots \\ \dot{d}_n \end{bmatrix} = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} x(t) \tag{3.7}$$

$$y(t) = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} \tag{3.8}$$

The complete explanation on how to derive the state space representation of an FDN can be found in (Smith, 2010). Here we attempt to illustrate it with an example. Consider the following simple FDN in figure 3.1. The total delay length for this FDN is $3 + 5 = 8$, hence the size of vector $\boldsymbol{u}$ is 8. We can express $\boldsymbol{u}$ and the state space matrices

as,

$$\boldsymbol{u} = \begin{bmatrix} d_{11} \\ d_{12} \\ d_{13} \\ d_{21} \\ d_{22} \\ d_{23} \\ d_{24} \\ d_{25} \end{bmatrix} \boldsymbol{A} = \begin{bmatrix} 0 & 0 & a & 0 & 0 & 0 & 0 & b \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & c & 0 & 0 & 0 & 0 & d \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \boldsymbol{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \boldsymbol{C} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{3.9}$$

The first 3 by 3 submatrix in the upper left corner of $\boldsymbol{A}$ represents the first delay line, and the 5 by 5 submatrix in the lower right corner of $\boldsymbol{A}$ represents the second delay line. Rows of $\boldsymbol{A}$ indicates the mixing destination state, and columns of A indicates the source of state mixing. However, we can only mix from the last state of each delay line, e.g. state d_{13} and d_{25}, to the first state of each delay line, e.g. state d_{11} and state d_{21}. Therefore a, b, c, d is only found at the third and eighth (3+5) element of the first and fourth (3+1) rows of A respectively. Since we would like to mix b portion of the second delay line to the first delay line, the constant b is placed at A_{18}. Similarly, since we want to mix c portion of the first delay line, third state, to the second delay line, first state, the constant c is placed at A_{43}.

Finally, the transfer function of the system can be found by,

$$H(z) = \frac{Y(z)}{X(z)} = \boldsymbol{C}(z\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{B} + D. \tag{3.10}$$

We can plot the frequency response by transforming frequency in radians θ to z: $z = e^{(} - i\theta)$, and plot Equation 3.10 for $0 \leq \theta \leq \theta$. If one would like to use frequency in Hz, then $z = e^{i(2\pi f)/f_s}$, where f_S is sampling frequency and $0 \leq f \leq f_s/2$.

Solving Equation 3.10 requires a lot of computational power when the order of the system is large. For a room with reverberation time of $0.5s$ and sampling rate of 44.1 kHz, it is required that total length of the delay lines (which is also the order of the FDN) M to be larger than 3308. Hence this require matrix inversion of size 3308×3308.

We can approximate the transfer function of the FDN with long delay lines by replacing them with shorter delay lines and keeping the relationship between delay lines constant, e.g: replace delay lines of length $\{20, 80\}$ with $\{2, 8\}$. The frequency response of the delay lines with shorter length is a stretched, and scale-down version of those with longer length. Figure 3.2 illustrates exactly that. Scaling down the FDN however does not work if the delay lengths are mutually prime.

Obtaining FDN frequency response from its transfer function is mathematically robust. However it comes at a huge computational cost to compute the matrix inversion in Equation 3.10. Hence we arrive at the second method to obtain the frequency response of an FDN, which is to run an impulse into the lossless FDN and record its output for a few seconds. Afterwards, we take the Fast Fourier Transform (FFT) of the output to obtain its frequency response. This method is much faster to implement although it is prone to round-off errors during FFT. These round-off errors however become negligible when longer signal is used at the expense of computation time.

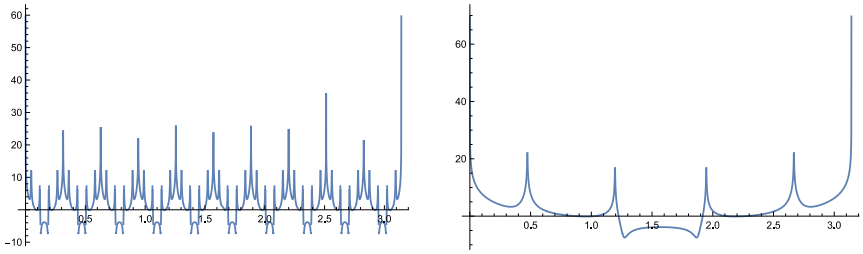

FIGURE 3.2: Left: Frequency response of original FDN with delay lengths 20 and 80. Right: Frequency response of scaled-down FDN with delay lengths 2 and 8. Size 2×2 Hadamard mixing matrix is used.

3.3.3 Ray-Tracing Delay Lines

In Chapter 1, we established the physical significance of signals inside FDN to model room acoustics. We set the lengths these delay lines is to set their values based on the time taken for each ray to reach the listener from the source after first order reflections. We call this *ray-tracing delay lines*. Later on in Chapter 4, we used these ray-tracing delay lines in our auralization algorithm. In the next section, we evaluate whether ray-tracing delay lines noticeably affect the colour of the lossless FDN output.

3.4 PROCEDURE

We test the ray-tracing delay lines in 15 room settings with various FDN sizes: 16, 32, 64, and 128. For details of the BRIRs, refer to the evaluation section of Chapter 6. The mixing matrix that we are going to use is the Hadamard matrix, since not only that it is maximally diffusive (Smith, 2010) but it is also computationally efficient when implemented using the Fast Hadamard Transform [3].

The famous echoic memory [4] study by Guttman and Julesz in 1963 found that it is difficult to detect more than 2s periodicity in Gaussian noise (Guttman and Julesz, 1963). However a more recent study in 2001 (Kaernbach, 2001) found that it takes at least 2.8s of periodicity in order to be unnoticeable by untrained participants. Hence for each method and FDN size, we run an impulse through the lossless FDN and obtain 3 seconds of its output. Any periodicity after 3s due to the lengths of delay lines is less likely to be noticeable. Therefore we obtained 131072 samples (44100 sampling rate) out from the lossless FDN using each BRIR setting. To compute the overall SFM, we took windows of 2048 samples and compute the SFM for each window, and finally obtain the average SFM. We report the overall SFM values in the next section.

[3] The Fast Hadamard Transform does not require any multiplicative operation, which is known to be more computationally demanding than addition or subtraction.

[4] Short-term memory for auditory information.

3.5 RESULTS

Table 3.1 contains the overall SFM values from all 15 BRIR settings and FDN sizes of $N = 16, 32, 64$, and 128. The SFM values generally increases as the number of delay lines used are increased, regardless of the room. This shows that the output of the loss FDN gets whiter as more ray tracing delay lines are used. The standard deviation (denoted as σ) also decreases as N increases, suggesting that the whiteness value of the lossless FDN output becomes stabler as N increases. From table 3.1, all values are above 0.3288, which is the JND amount found in Chapter 4 in terms of SFM. This shows that the colouration caused by these ray-tracing delay lines are most likely not noticeable.

TABLE 3.1: Mean SFM values from 2048 windows out of 131072 samples in total

BRIR	16	32	64	128
R1 P1	0.433	0.536	0.547	0.552
R1 P2	0.474	0.519	0.536	0.533
R1 P3	0.494	0.541	0.554	0.546
R1 P4	0.544	0.532	0.540	0.542
R2 P1	0.460	0.557	0.561	0.555
R2 P2	0.463	0.562	0.564	0.558
R2 P3	0.418	0.557	0.558	0.554
R3 P1	0.454	0.516	0.535	0.525
R3 P2	0.404	0.531	0.542	0.547
R3 P3	0.413	0.515	0.539	0.540
R3 P4	0.505	0.556	0.559	0.556
R4	0.585	0.576	0.571	0.564
R5	0.622	0.585	0.571	0.567
R6 P1	0.504	0.560	0.560	0.556
R6 P2	0.488	0.556	0.559	0.559
μ	0.484	0.547	0.553	0.550
σ	0.062	0.021	0.012	0.012

3.6 CONCLUSION

In this chapter we explained how the output of a lossless FDN is affected by its delay lengths, and whether setting the delay lengths of the FDN using ray-tracing delay line lengths setting is feasible. Ideally, when one feeds an impulse to a lossless FDN, the output should resembles that of a white noise, meaning that it neither boost nor cut any frequency component. Theoretically, it is possible to obtain the transfer function of the FDN and plot its frequency response. However in practice, N (the size of FDN) is

large and the delay lengths can grow up to thousands of samples in length. Therefore it may be practically impossible or too tedious to obtain its frequency response.

We can then obtain its frequency response by running an impulse through the FDN and record its lossless output for at least 3 seconds (131072 samples, 44100 sampling rate in our case), and quantify the amount of whiteness in that output using well known whiteness measures such as the SFM or Q-value as introduced in Chapter 2. Our results show that ray-tracing delay lengths setting do not introduce coloration to the FDN beyond that of noticeable amount established in Chapter 4. In other words, by using ray-tracing delay lines, the output of the lossless FDN is still potentially perceived as being white. This shows that it is reasonable to establish the physical significance of the signal in the FDN and set the delay lengths based on the paths of the first order reflections. In the next chapter, we proceed by introducing our algorithm for binaural room auralization, which used the ray-tracing delay lines.

Chapter 4

Minimally Simple Binaural Room Modelling Using a Single Feedback Delay Network

This work is based on the peer-reviewed manuscript: Agus, N., Anderson, H., Chen, J.M., Lui, S., "Minimally Simple Binaural Room Modelling Using a Single Feedback Delay Network," Journal of the Audio Engineering Society (Jun 2018). Manuscript accepted (with editor).

4.1 ABSTRACT

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. This requires modeling up to third order reflections and generalize the higher reflection orders. We would like to further reduce the computational cost by explicitly only modeling the first order reflections, and gracefully degrade the accuracy of higher orders instead of generalizing the entire higher orders. In order to do this, in the previous chapters we introduced the physical significance of audio signals in the FDN, which leads to the ray-tracing delay lines, as well as basic mathematical frameworks for our room acoustic modeling algorithm. In Chapter 3, we also showed that ray-tracing delay lines do not introduce coloration to the FDN beyond that of noticeable level (JND), which was established in Chapter 2. Therefore now 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.

4.2 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 (Välimäki et al., 2012).

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) (Allen and Berkley, 1979), 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 (Savioja and Svensson, 2015). 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 (Pelzer et al., 2014).

Applications of both categories of reverberation algorithms include acoustic simulation for training simulations, 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) (Siltanen et al., 2007) and a Feedback Delay Network (FDN) (Jot and Chaigne, 1991) with a bank of head related transfer function

filters (HRTF) (Duda, Algazi, and Thompson, 2002) and a bank of interaural time difference (ITD) delay lines to simulate binaural room acoustics in real time.

4.3 RELATED WORK

Many acoustic modeling systems work by pre-computing 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 (Välimäki et al., 2012; Välimäki et al., 2016).

This category of hybrid GA approaches includes methods that range from simple auralization algorithms to extensive room modeling systems such as DIVA (Savioja et al., 1999; Savioja, Lokki, and Huopaniemi, 2002) and RAVEN (Schröder, 2011; Schroder and Vorlander, 2007). 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 (Savioja et al., 1999). 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 (Gerzon, 1973). 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 (Schröder, 2011). 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 (Schröder, Dross, and Vorländer, 2007). 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 (Schröder and Lentz, 2006). RAVEN updates its early reverb simulation more frequently than the late reverb.

In (Menzer, 2012), 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 (Menzer, 2012) 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 (Menzer and Faller, 2009). The work in (Menzer and Faller, 2009) 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 (Menzer, 2010). 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 (Menzer and Faller, 2009), 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 1st 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 (Wendt, Par, and Ewert, 2014a). 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 (Iso3382-1, 2009). The listening test shows that the method has good perceptual accuracy, compared to the measured BRIRs. However this method in (Wendt, Par, and Ewert, 2014a) 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 (Bai, Richard, and Daudet, 2015), 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 (Savioja et al., 1999). 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 (Wendt, Par, and Ewert, 2014a; Menzer, 2012; Bai, Richard, and Daudet, 2015; Savioja et al., 1999) 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 (Bai, Richard, and Daudet, 2015). The most significant difference between the method presented here and the one in (Bai, Richard, and Daudet, 2015) is described in section 4.4 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.

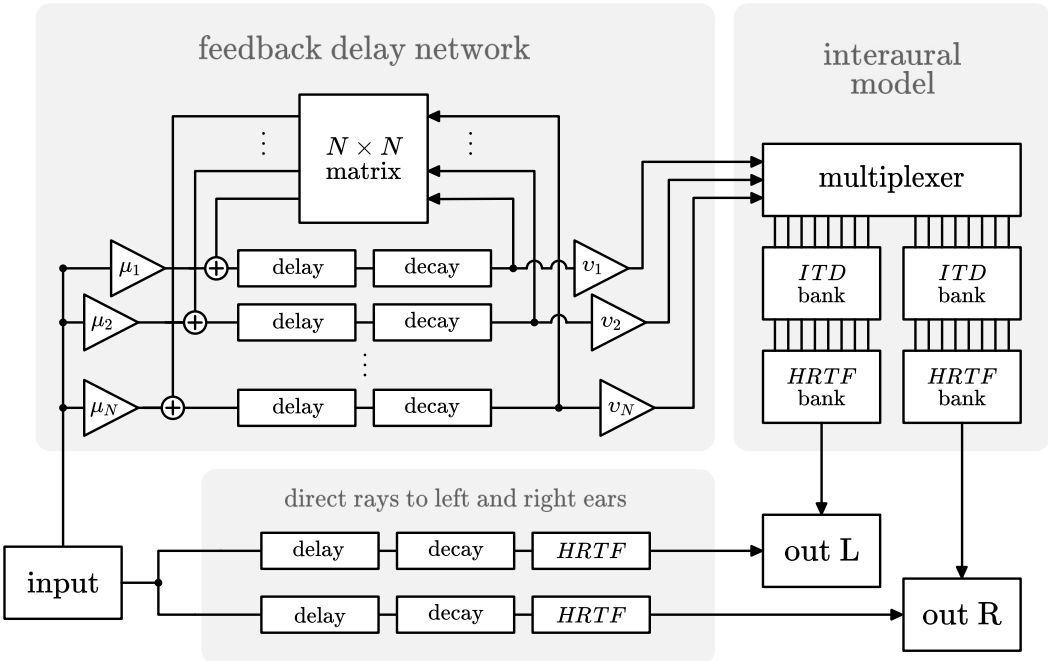

FIGURE 4.1: The proposed system: the delay lengths and output gain coefficients in the FDN are chosen so that the first impulses to come out from the network are the early reflections as modeled by the Acoustic Rendering Equation. Each delay line in the network corresponds to one patch of surface geometry in a 3D model of an acoustic space. By setting appropriate gain coefficients μ_n at the input and v_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.

4.4 METHOD

4.4.1 Method Overview

Figure 4.1 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 v 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 (Griesinger, 1996) 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.

4.4.2 The Acoustic Rendering Equation

Our model of 1st order reflections is a standard application of the Acoustic Rendering Equation (ARE) (Siltanen et al., 2007). The ARE was first proposed by Siltanen et al. in (Siltanen et al., 2007). The physics and language underlying the ARE are based on radiometry and optics. For readers who are unfamiliar with those topics, we give a derivation of the ARE starting from first principles of acoustics in (Agus et al., 2017). We refer readers to our previous work in (Anderson et al., 2017) (detailed in Chapter 4) for further details on how we model the first order reflections using the ARE. In this chapter, we use the same notations as that in our previous work in (Anderson et al., 2017).

For the purpose of clarity, we reiterate that Siltanen et al. define the ARE as follows,

$$\ell\left(\boldsymbol{x},\Omega\right)=\ell_0\left(\boldsymbol{x},\Omega\right)+\int_{\mathcal{G}}R\left(\boldsymbol{u},\boldsymbol{x},\Omega\right)\ell\left(\boldsymbol{u},\frac{\boldsymbol{x}-\boldsymbol{u}}{|\boldsymbol{x}-\boldsymbol{u}|}\right)\mathrm{d}\boldsymbol{u}, \tag{4.1}$$

where ℓ is the total outgoing radiance and ℓ_0 is the emitted radiance at $\boldsymbol{x}$ to direction Ω. The integral term represents reflected radiance from all other points $\boldsymbol{u}$ in the room.

To simplify the notations, as stated previously in Chapter 4, we define $\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}$, a unit vector pointing in the direction from $\boldsymbol{u}$ to $\boldsymbol{x}$,

$$\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}=\frac{\boldsymbol{x}-\boldsymbol{u}}{\|\boldsymbol{x}-\boldsymbol{u}\|}. \tag{4.2}$$

Using this notation we rewrite the ARE as follows,

$$\ell\left(\boldsymbol{x},\Omega\right)=\ell_0\left(\boldsymbol{x},\Omega\right)+\int_{\mathcal{G}}R\left(\Lambda_{[\boldsymbol{u},\boldsymbol{x}]},\boldsymbol{x},\Omega\right)\ell\left(\boldsymbol{u},\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}\right)\mathrm{d}\boldsymbol{u}. \tag{4.3}$$

The Neumann series solution of equation (5.4) is accordingly,

$$\ell_{n+1}(\boldsymbol{x},\Omega)=\int_{\mathcal{G}}R\left(\Lambda_{[\boldsymbol{u},\boldsymbol{x}]},\boldsymbol{x},\Omega\right)\ell_n\left(\boldsymbol{u},\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}\right)\mathrm{d}\boldsymbol{u}. \tag{4.4}$$

In our implementation, we discretise the surface geometry $\mathcal{G}$ into a set of discrete patches $A_n\subset\mathcal{G},$ for $n=1...N$ and use Monte-Carlo integration to compute the integral in equation (5.4) for each patch.

4.4.3 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 nth surface patch in the 3D model,

$$E(A_n, L) = \left(\frac{N\ \Phi(F_n)}{\pi \mathcal{G}}\right) \int_{A_n} h(\boldsymbol{x}, L)\ \mathrm{d}\boldsymbol{x}. \tag{4.5}$$

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

Equation (4.5) is based on the assumption that as time progresses the reverberated energy increasingly approaches an evenly diffused and mixed state (Griesinger, 1996). 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 (4.5) 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 (Anderson et al., 2017).

$h(\boldsymbol{x}, L)$ is the point collection function, similar to what is defined in our previous work (Anderson et al., 2017), with the addition of the absorption term ξ,

$$h(\boldsymbol{x}, L) = \xi(\boldsymbol{x}, L)\ \mathcal{V}(\boldsymbol{x}, L)\ P(\boldsymbol{x}, L). \tag{4.6}$$

The absorption ξ and visibility $\mathcal{V}$ terms are defined as in (Siltanen et al., 2007). The geometry term $P(\boldsymbol{x}, L)$ is also defined as in (Anderson et al., 2017).

The constant N in (4.5) 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 the block-circulant mixing matrix proposed in (Anderson et al., 2015), which requires

[1] We must select an appropriate mixing matrix to ensure that this is true. One example is the Hadamard matrix (Jot, 1997)

only that N be a multiple of some integer K, but needs more time to reach an evenly mixed state when K is small.

4.4.4 Gain Coefficients υ_n at the FDN Output

Let F_n denote the output of the nth 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,

$$F_n{}^2 = \Phi(F_n). \tag{4.7}$$

We define the gain coefficient υ_n as follows,

$$\upsilon_n = \sqrt{\frac{N}{\pi \mathcal{G}} \int_{A_n} h(\boldsymbol{x}, L) \mathrm{d}\boldsymbol{x}}. \tag{4.8}$$

We can confirm by inspection that the following relation holds,

$$E(A_n, L) = (F_n \upsilon_n)^2. \tag{4.9}$$

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

4.4.5 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 1st order reflection using the ARE. Afterwards, we need to collect that energy at the listener position.

In equation (4.10) below, $E_n(A_n, L)$ denotes the *acoustic irradiance* at the listener position L due to 1st order emitted radiance ℓ_1 at A_n, the nth surface patch in our 3D model. Irradiance is a measure of incident energy flux per unit area,

$$E_1(A_n, L) = \int_{A_n} h(\boldsymbol{x}, L) \ell_1(\boldsymbol{x}, \Omega) \mathrm{d}\boldsymbol{x}. \tag{4.10}$$

4.4.6 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 1st order reflection off the nth surface patch.

Recall from equation (4.10) that $E_1(A_n, L)$ denotes the irradiance at the listener due to 1st order reflections off the patch A_n. It follows that the following relation must hold,

$$E_1(A_n, L) = (F_n \beta_n)^2. \tag{4.11}$$

The term ℓ_1 in equation (4.10) can be computed by applying the ARE in the usual way[2].

However, directly multiplying the input or output of the nth 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 1st order reflection gain. Instead, we define μ_n to be a gain coefficient at the input of the nth channel of the FDN and set it as follows,

$$\mu_n = \beta_n / \upsilon_n. \tag{4.12}$$

The effect of this is that for 1st order reflections, the output coefficient υ_n is canceled out by the input and the resulting 1st order FDN output is exactly as given in equation 4.11, but for second and higher order FDN output, the result is as specified by equation 4.9, 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 4.1 for a diagrammatic representation of this.

4.4.7 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 4.1 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 (Brown and Duda, 1998). 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

[2]We refer readers unfamiliar with this method to (Agus et al., 2017), where the authors explain it in detail.

between 2 to 7 kHz that is typically caused by ear canal and concha resonance (Shaw and Teranishi, 1968) and the high frequency roll-off or notch above 8kHz depending on front-back configuration (Carlile, 2013). Further explanations can also be found in (Ballachanda, 1997; Hellstrom and Axelsson, 1993). While it is impossible to model every individual HRTF, one may add simple pole-zero EQ and 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.

4.4.8 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 4.1 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 nth delay line to correspond to the timing of the first-order reflection that comes from the nth 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 4.1 using equation (5.37).

4. Using the Acoustic Rendering Equation in (Siltanen et al., 2007), model the 1st order reflections and compute the amount of energy at the listener due to 1st order reflections using equation (4.10). 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 1st order reflection. Compute it with equation (4.11) using values from (4.11) obtained in the previous step. Then the coefficient μ_n at the input of the nth 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 nth patch of surface geometry that we estimated in step 1 above.

4.5 OBJECTIVE EVALUATION

4.5.1 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 (Jeub, Schäfe, and Var, 2009) 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 entry-way 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 4.5.5. The rest of the configurations were used for listening test explained in section 4.6.2 instead. For BRIRs from (Jeub, Schäfe, and Var, 2009), 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 (Farina, 2000). 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 (Hak, Wenmaekers, and Luxemburg, 2012). The response of the speaker is shown in Figure 4.2. 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 (Lundeby et al., 1995) 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.

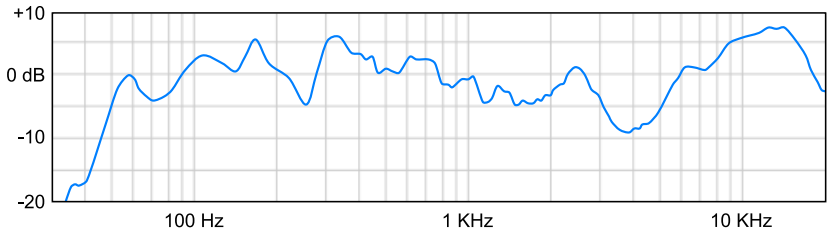

FIGURE 4.2: The frequency response of the omni-directional speaker used to measure BRIRs in R1, R2, and R3.

4.5.2 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 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 (Smith, 2010).

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 (Allen and Berkley, 1979) 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 (Wendt, Par, and Ewert, 2014a).

2. *Baseline Method 2 (Baseline ARE)*: We generate the the BRIR up to second order reflections using the ARE (Siltanen et al., 2007) 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 4.4.7).

4.5.3 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 4.1. 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 4.1 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 (Allen and Berkley, 1979). 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 4.1 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 (Lentz, 2007) 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 (Brungart, Simpson, and Kordik, 2005), 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 4.1, 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 4.1: 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.

4.5.4 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 (Iso3382-1, 2009). 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 (Hidaka, Beranek, and Okano, 1995) 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 (Fechner, 1966). 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.

4.5.5 Results

Comparison with Measured BRIR

BRIR	Meas	Prop Method	Meas.	Prop Method	Meas.	Prop Method	Meas.	Prop Method	Mea.	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 4.2: The values of all five room parameters of the measured BRIRs (Meas.) and simulated BRIRs using the proposed method (Prop. Method) with 128 patches. Results in bold are more than 1 JND from the measured result.

Table 4.2 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 (Wendt, Par, and Ewert, 2014a) 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 (Vigeant et al., 2015) 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 (Iso3382-1, 2009), which we are using to report the data in Table 4.2. 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 4.2 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 (Iso3382-1, 2009). 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 4.2, a significantly higher EDT error is observed in R6. The proposed model actually may even actually yield higher error with a more detailed 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 (Jeub, Schäfe, and Var, 2009).

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 4.3: 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 4.5.2, which use a separate multi-tap delay and FDN for early reflections and late reverb. One baseline method simulates early reflections up to the 3rd order using the image source method and the other uses the acoustic rendering equation up to the 2nd 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 4.5 and 4.6 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 4.3, 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

Baseline Method	Mesh Size 32 ISM	Mesh Size 32 ARE	Mesh Size 64 ISM	Mesh Size 64 ARE	Mesh Size 128 ISM	Mesh Size 128 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 4.4: 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.

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 (Wang, Rathsam, and Ryherd, 2004) 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 (Wendt, Par, and Ewert, 2014a) and (Tenenbaum et al., 2007) 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 1st 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 4.1 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 4.4 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 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.

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 4.5: 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 4.6: 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.

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.

4.6 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 (Lindau, 2015). The procedure and result is presented in sections 4.6.1 and 4.6.1, 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 4.6.2 and 4.6.2 respectively.

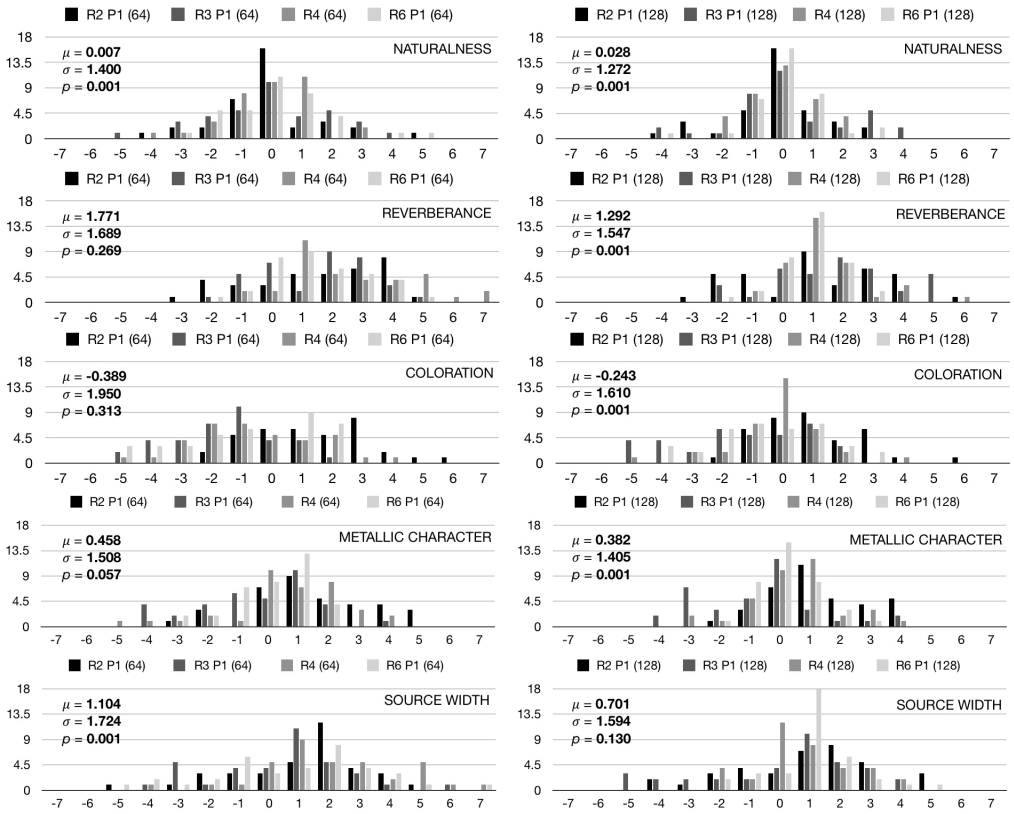

FIGURE 4.3: Histogram of the 15-scale bipolar ratings by 19 subjects on all five perceptual qualities, using the synthesized signal from the proposed method with 64 patches (left) and 128 patches (right). The rating scale is explained in section 4.6.1. The mean (μ) and standard deviation (σ) of the rating 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.

4.6.1 Part I: Listening Test Evaluation of Standard Perceptual Qualities

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 4.5. 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 (Siltanen, Lokki, and Savioja, 2010), we filtered the dry audio signals to exclude frequencies below 100 Hz. To ensure fair comparison across listening test subjects of various ages, we also filtered out frequencies above 15 kHz as recommended in (Stelmachowicz et al., 1989).

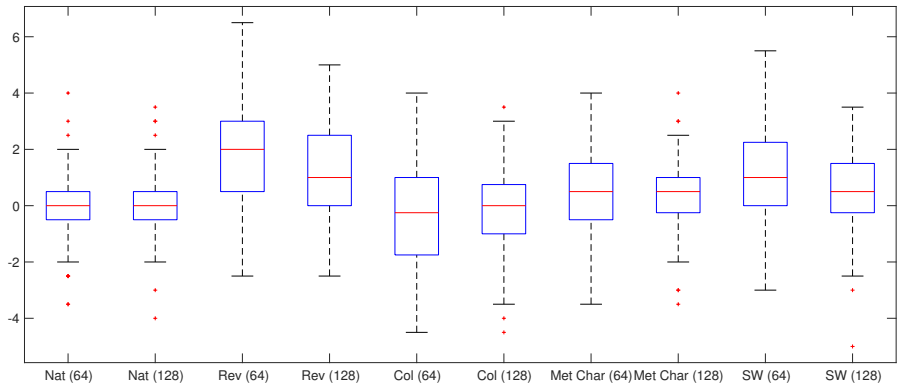

FIGURE 4.4: Boxplots of the 15-scale bipolar ratings by 19 subjects on all five perceptual qualities: Naturalness (Nat), Reverberance (Rev), Coloration (Col), Metallic Character (Met Char), and Source Width (SW) using 64 and 128 patches. Each boxplot contains 76 responses in total from 19 subjects and 4 different room condition.

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 (Chaiken and Eagly, 1983; South, Oltmanns, and Turkheimer, 2005).

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.

Results

Figure 4.3 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 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 4.4 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.

4.6.2 Part II: Measuring the Sense of Spatial Location

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 (Woirgard et al., 2012). 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 4.5 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.

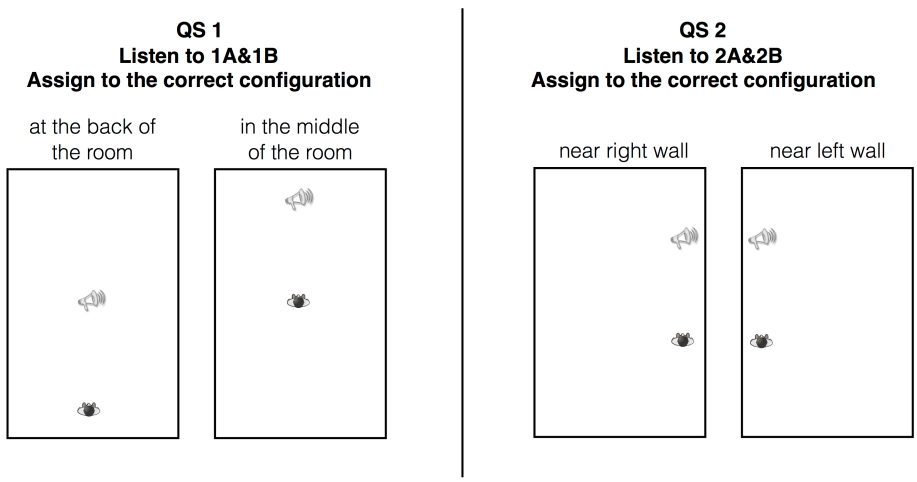

FIGURE 4.5: Sample listening test question

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 4.5 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.

Result

Figure 4.6 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).

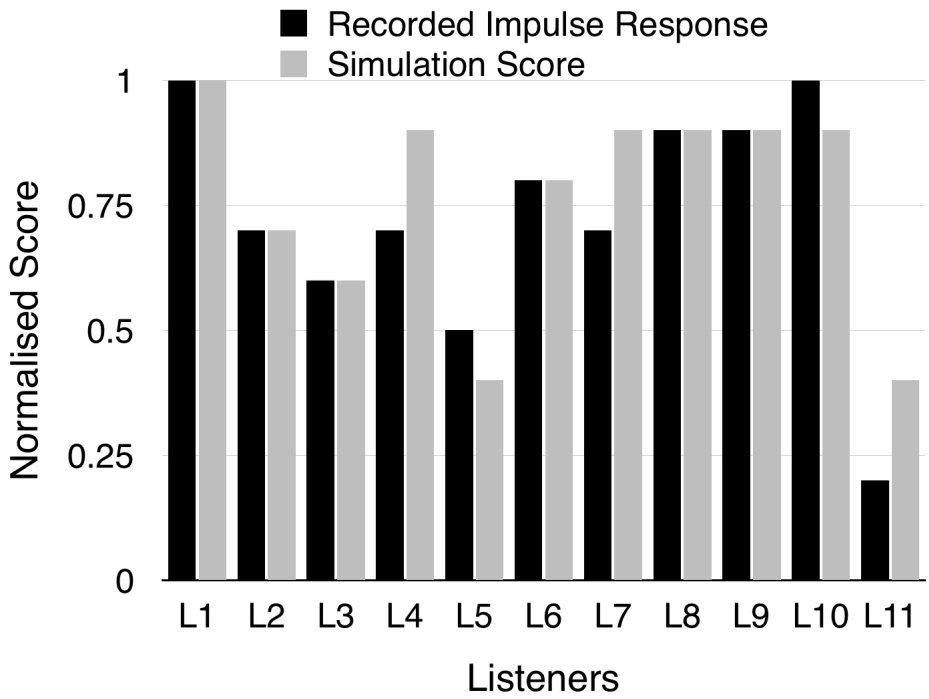

FIGURE 4.6: Normalised listening test scores of 11 test participants, comparing results for measured (black) and simulated (grey) reverb impulse responses.

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 (Harris and Holland, 2009), 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.

4.6.3 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 4.5) 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.

4.7 CONCLUSION

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.

4.8 FUTURE WORK

Over the course of this work 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 (Anderson et al., 2017), 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 4.5 and observed significant improvements. We expect that similar improvements are possible with many other related methods.

Chapter 5

Modeling the Proportion of Early and Late Energy in Two-Stage Reverberators

This work is based on the peer-reviewed manuscript: Anderson, H., Agus, N.*, Chen, J.M, Lui, S., "Modeling the proportion of early and late energy in two-stage reverberators," Journal of the Audio Engineering Society, Vol 65(12), 1071-1031, (Dec 2017).* [1]

5.1 ABSTRACT

In the previous Chapter (Chapter 4) we presented our hybrid binaural room acoustic auralization algorithm. We used principles that we derived in Chapter 1 in our algorithm. The main difference between our work in Chapter 4 and other related works in the literature (hybrid geometrical acoustic modeling) is that in our model, both first order and higher order reflections are modeled, but the accuracy of the higher orders are gracefully degraded as the order grows higher, while in other hybrid models, the higher orders are generalized and is often assumed to be constant, i.e: independent of the room geometry, materials, source, or listener locations. However this assumption is not always the case.

The following is one example. Clarity Index, denoted by C_{80}, is a unit of measurement that quantifies the ratio of early to late reverb energy on a log scale. We measure a standard deviation for Clarity Index of more than eight decibels across various listener positions in large rooms, indicating that it varies audibly with respect to location. The most efficient acoustic modeling reverberators use a two stage model, producing detailed early reflections with generic late reflections. Most methods of this type do not accurately model energy flux through the late reverb module, hence their Clarity Index is inaccurate. In this chapter, we propose an efficient method to model late reverb energy flux based on principles we established in the previous chapters. Our method proposed in this chapter can be applied to existing hybrid room acoustic algorithms, i.e: method proposed in this chapter is the modification of the methods introduced in Chapter 1 and 4 such that it can be applied to existing hybrid algorithms. We show that it models clarity and the related metrics, Definition (D_{50}) and Centre Time (T_S), more than twice as accurately as the baseline method.

[1] (*) Both authors contributed equally.

5.2 INTRODUCTION

Computer simulations of reverberant room acoustics have applications in music recording, video games, research, and movie production. They are also used to render realistic audio-visual scenes for various testing, training, and rehabilitation exercises. The required level of accuracy in a reverb simulation depends on the application. In gaming applications, any perceptually plausible result is sufficient but when using a virtual enviornment to investigate the effect of aircraft flyover noise we require more precise modelling (Arntzen, Bertsch, and Simons, 2015; Rizzi, 2013).

The methods to simulate room acoustics are generally divided into two broad categories, Numerical Acoustics (NA) and Geometrical Acoustics (GA). Simulation results using NA is more accurate than GA, as it involves solving the wave equation in three dimension. Wave phenomena such as diffraction and interference can be precisely captured in the solution. However, methods that involve solving the wave equation for the entire duration of impulse response are likely too computationally intensive (Välimäki et al., 2012), especially for use cases where we want to generate new reverb impulses in real time as the listener moves about the room. Several systems have been proposed that achieve more accurate acoustic simulation by pre-computing impulse responses for a large number of listener and sound-source locations. However the time required to compute impulse responses for a large number of source and listener positions may range from hours to days, depending on the method. Storage space requirements, along with the complexity of implementation, form a barrier to their being widely adopted.

In GA methods, sound waves are treated as rays. Therefore, low-frequency wave phenomena such as diffraction and interference is not captured in GA methods. However it is still a plausible approach especially at the frequencies where the wavelength is relatively small to the room dimension (Välimäki et al., 2012). GA methods are much faster to run than NA methods, but at the cost of accuracy. Also, they typically do not compute the higher order reflections explicitly because the complexity of reflections models, such as the Image Source Method (ISM) (Allen and Berkley, 1979), grows exponentially as we increase the maximum order of reflections in the model. An exact simulation of late reverberation is not feasible. Most acoustic modeling methods that require only perceptual plausibility and not a strict numerical accuracy strive for accuracy in the early reflections because the early part of the impulse response is considered to be more perceptually relevant (Lehmann and Johansson, 2010). These methods combine a precise early reflections model with a light-weight approximation of late reverberation.

The most efficient examples of this idea allow the listener and sound source location to change in real time without using pre-computed impulse responses (Välimäki et al., 2012). This is a very desirable feature for applications where perceptual measures take precedence over numerical accuracy. One such example is the DIVA auralization system, where early reflections are calculated using the ISM or beam tracing method, and late reverberation is handled by an FDN-type reverberation structure (Savioja et al., 1999).

	P1	P2	P3	P4	P5	P6	P7	σ	JND	$\sigma/(1\text{ JND})$
C_{80} (dB)	10.8	5.88	3.32	-0.29	-3.5	-9.79	-13.03	8.50	1	8.50
D_{50}	0.92	0.77	0.64	0.40	0.22	0.07	0.0	0.35	0.05	7.11
T_S (ms)	547	228	339	526	734	936	943	279	10	27.9

(A) C_{80}, D_{50}, and T_S of measured impulse responses in Aula Carolina Cathedral ($5700m^3$), with RT_{60} of 4.7s and EDT of 3.2s, averaged in the 500Hz and 1000Hz octave bands.

	P1	P2	P3	P4	P5	σ	JND	$\sigma/(1\text{ JND})$
C_{80} (dB)	6.69	5.76	4.23	2.59	3.16	1.71	1	1.71
D_{50}	0.73	0.63	0.50	0.36	0.35	0.31	0.05	3.43
T_S (ms)	43.8	56.0	70.3	85.1	83.7	88.0	10	1.80

(B) C_{80}, D_{50}, and T_S of measured impulse responses in Aachen Lecture Hall ($412m^3$), with RT_{60} of 0.96s, and EDT of 0.85s, averaged in the 500Hz and 1000Hz octave bands.

	P1	P2	P3	σ	JND	$\sigma/(1\text{ JND})$
C_{80} (dB)	12.09	8.99	10.31	1.55	1	1.55
D_{50}	0.89	0.74	0.68	0.11	0.05	2.16
T_S (ms)	21.40	37.35	43.78	11.52	10	1.15

(C) C_{80}, D_{50}, and T_S of measured impulse responses in Aachen Office Room ($92.8m^3$), with RT_{60} of 0.56s and EDT of 0.49s, averaged in the 500Hz and 1000Hz octave bands.

	P1	P2	P3	P4	P5	σ	JND	$\sigma/(1\text{ JND})$
C_{80} (dB)	18.48	16.51	15.11	16.91	17.47	1.24	1	1.24
D_{50}	0.95	0.93	0.92	0.94	0.94	0.01	0.05	0.22
T_S (ms)	16.05	17.90	22.30	18.60	18.15	2.29	10	0.23

(D) C_{80}, D_{50}, and T_S of measured impulse responses in Aachen Meeting Room ($124m^3$), with RT_{60} of 0.37 and EDT of 0.21s, averaged in the 500Hz and 1000Hz octave bands.

TABLE 5.1: C_{80} (dB), D_{50}, and T_S (ms) values averaged for the 500Hz and 1000Hz octave bands at various source-microphone positions in a large cathedral and small lecture room. We obtained the impulse responses from (Jeub, Schäfe, and Var, 2009). Note that the standard deviation, σ, of the values taken across several points in the same room is greater than 1 JND in the majority of the cases. Also, the variation is much more extreme in rooms that have higher reverberation time.

5.2.1 Is the variation of Clarity Index with respect to position an audible effect?

The acoustic units that quantify the proportion of early energy to late energy are definition (D_{50}), clarity index (C_{80}), and center time (T_S) (Iso3382-1, 2009). D_{50} is defined as the ratio between the energy in the first 50ms to the total energy of the impulse response. C_{80} is the ratio between the energy of the first 80ms to the late energy of the impulse response. The third parameter, T_S, is the center of gravity of the energy of the impulse response. The Just Noticeable Difference (JND) values for D_{50}, C_{80}, and T_S are 0.05, 1 dB, and 0.01 seconds of the arithmetic mean in the 500Hz and 1000Hz frequency bands respectively (Iso3382-1, 2009).

To quantify the importance of modeling early to late energy balance, we calculated the C_{80}, D_{50}, and T_S for impulse responses taken at various listener and source positions in several rooms. The results shown in table 5.1 are examples measured from publicly-available impulse responses recorded in four different rooms: a cathedral called Aula Carolina ($5700m^3$), in a medium-sized room called Aachen Lecture Hall ($412m^3$), and two relatively small rooms called Aachen office ($92.8m^3$) and meeting room ($124m^3$) which are part of the Aachen Impulse Response (AIR) database (Jeub, Schäfe, and Var, 2009). Although there are more room impulse responses used in section 5.5, we present responses only from these rooms in table 5.1 as they are the only ones with a multiple source-microphone configurations.

In table 5.1 we see that the variation due to changes in listener and source position in the Aula Carolina cathedral is extreme but in the smaller lecture hall it is only slightly more than one Just Noticeable Difference (JND). Our informal experiments indicate that the variance of C_{80}, D_{50}, and T_S increases with the reverberation time of the room. The same trend can be observed in table 5.1. It can be less than one JND in rooms that have very short reverberation time, such as the meeting room (0.37s).

5.3 Related Work

The proposed method applies to the simplest and most efficient class of two-stage acoustic modeling reverberators, using delay networks that appear to be most widely used due to their computational efficiency and perceptual plausibility (Välimäki et al., 2012). Figure 5.1 shows a typical example. Existing methods of this type are proposed in several previous works (Jot, 1997; Menzer, 2010; Wendt, Par, and Ewert, 2014a; Wendt, Par, and Ewert, 2014b; Carty and Lazzarini, 2010; Sarti and Tubaro, 2001). We will discuss these in more detail in this section. We will also discuss an example where this hybrid method is used in more extensive auralisation programs (Savioja et al., 1999; Savioja, Lokki, and Huopaniemi, 2002).

Jot proposes an early design for an efficient two stage reverberator (Jot, 1997) that used a multi-tap delay connected in series to a FDN. The multi-tap delay produces early reflections and also provides input diffusion for the FDN to increase its echo density. This idea is based on the reasoning that the early reflections are the most perceptually significant part of the reverb impulse response and therefore we can create perceptually plausible reverberation by coupling detailed early reflections to generic late reverb.

We observe the same concept applied in (Carty and Lazzarini, 2010), where the authors propose a lightweight implementation of a hybrid reverberation algorithm using Csound opcodes. The purpose of this system is to produce natural sounding reverberation for use in binaural room simulations using head related transfer functions (HRTFs).

Another example is found in (Savioja et al., 1999; Savioja, Lokki, and Huopaniemi, 2002) where the authors describe DIVA, an extensive virtual acoustic space design that can be used to artificially render 3D sound of a given space in real-time. In both papers, it is shown that system connects the output from the early reflection to the late reverberation unit via several attenuation coefficients based on material parameters and also via an attenuation proportional to $1/r$, where r is the distance from each image source to the listener.

A similar structure is proposed in (Wendt, Par, and Ewert, 2014a; Wendt, Par, and Ewert, 2014b), where early reflections up to the Nth order are calculated using the Image Source Method and late reverberation is produced by a 12-channel FDN. The configuration of the FDN does not depend on the listener or source position. In the FDN of order 12, three sets of four delay lines each model the length, width, and height of the room and the delay lengths are set to represent those dimensions. The outputs of the highest order early reflections are the input to the FDN. These delay taps are attenuated to model energy dissipation due to spherical spreading of the wave-front during its propagation from image source to listener position. This attenuation is applied before the signals enter the FDN. Extensive objective evaluations are done in (Wendt, Par, and Ewert, 2014a) to show that there is good agreement between their simulation results and recorded binaural room impulse response (BRIR). However, in section 5.5, we show that improvements are evident after an attempt to balance early and late energy is made.

Another related reverberation algorithm variation is presented in (Sarti and Tubaro, 2001), where early reflections and late reverberation are simulated using the image source method and a waveguide digital network (WDN) respectively. The proposed design in (Sarti and Tubaro, 2001) models pure specular reflection. Like the other methods described above, the output of the highest order early reflections serves as an input to the WDN.

Menzer proposes a different reverberation structure that produces early reflection and late reverberation using two parallel $FDNs$ (Menzer, 2010). The first FDN produces first and second order reflections using parameters calculated from the image source method. The second FDN produces late reverberation that matches the interaural coherence and energy decay relief of a reference $BRIR$. Menzer states that only these two features are perceptually relevant for late reverberation (Menzer and Faller, 2009). In (Menzer, 2012), he presents an implementation of his method on a mobile device. It is capable of rendering reverb with different source and listener positions in a virtual room that is perceived as perceptually plausible to both trained and untrained listeners. The author did not consider the effect of changes in source and listener position when rendering the late reverberation, probably because of the storage requirements of saving pre-computed BRIRs for a suitably large number of combinations of source and listener positions in a room. Therefore the late reverberation energy is assumed to be the same for all listener and source positions in a room and the balance of

energy between early and late reflections in different listener-source configurations are not considered.

None of the methods mentioned above explicitly models C_{80}, D_{50}, and T_S because they send energy from the early reflections output into the late reverb without adjusting the gain to account for the changes in the proportion of early to late reflected energy with respect to listener and source position. The difficulty is that a physically relevant reflections model should have the late reverb taking its input from reflections off of the walls of the room. However the early reflections output represents acoustic energy incident on the listener position, not the walls. If we consider the question, "what proportion of the reflected energy in the room reaches the listener?", we see that for early reflections, the answer depends on the sound source position. If both sound source and listener are close to a wall then a relatively large proportion of the early reflections will reach the listener, compared to the situation where the source is far from the listener. However, for late reverb the situation is different. Because late reverb energy is diffused and mixed more evenly around the room (Griesinger, 1999), the proportion of late energy that reaches the listener depends mainly on the listener location; the source location is much less relevant. As a result, the balance of early and late reflected energy changes with respect to source position, listener position, and room geometry, producing changes in C_{80}, D_{50}, and T_S. If we want to model the late reverb input correctly, we need to know the energy level at the walls of the room.

There exist several methods that do calculate the energy level in both early and late reverb. However, we have not found any that can be adapted to existing efficient two-stage hybrid structures discussed in section 5.3.1.

The method we present here is an improvement that can be applied to any hybrid reverberation algorithm with structure similar to the one shown in figure 5.1. Our main idea is to estimate the amount of energy reflecting off the walls at the time of the highest order early reflections and adjust the gain of the signal connecting the early reflections unit to the late reverb unit so that the energy entering the late reverb unit corresponds to the acoustic simulation. This will allow us to model the C_{80}, D_{50}, and T_S while still maintaining a simple and efficient real time implementation.

The remaining parts of the paper are organized as follows. In section 5.4, we describe our three key contributions, which are to unify the measurement of energy in the early and late reverberator, calculate attenuation coefficients β_h for the early reflections mixing into the late reverb input, and υ for the late reverb output. Section 5.5 evaluates the efficacy of the method with respect to modeling C_{80}, D_{50}, and T_S . (**AgusEnergyBased**) provides supporting materials to clarify the mathematics in section 5.4.

Previous work that would benefit from applying our proposed method include, (Jot, 1997; Menzer, 2010; Wendt, Par, and Ewert, 2014a; Wendt, Par, and Ewert, 2014b; Carty and Lazzarini, 2010; Sarti and Tubaro, 2001; Savioja et al., 1999), and any other designs that combine detailed early reflections together with a generic late reverb structure.

5.3.1 Structure of Existing and Proposed Methods

Figure 5.1 is a block diagram of a typical example of an efficient two-stage acoustic modeling reverberator. The early reflections are produced by a multi-tap delay whose

output tap times and their gains are based on a detailed reflection model. The late reverb is produced by either a Feedback Delay Network (FDN), as shown in the figure, or by convolution with a pre-computed impulse response (not shown). The late reverb module may model some of aspects of the sound of the room. For example, it may model inter-aural correlation or reverb decay time. However, it is not an accurate and detailed model of individual reflections.

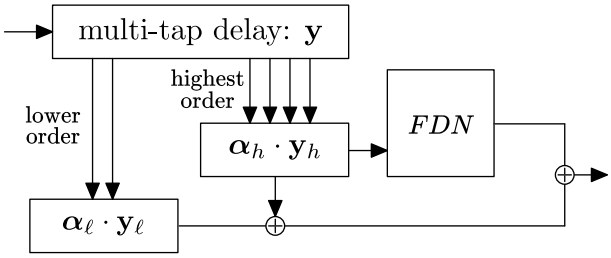

FIGURE 5.1: Typical structure of existing efficient two-stage hybrid acoustic models. A multi-tap delay generates early reflections and an FDN produces late reverb. For each input sample, the delay produces a vector of output samples $\mathbf{y}$. A vector of gain coefficients $\boldsymbol{\alpha}$ scales and mixes the elements of $\mathbf{y}$ by vector dot product. The scaling takes place in two parts, $\boldsymbol{\alpha}_l \cdot \mathbf{y}_l$ is the lower order reflections, which do not provide input to the FDN and $\boldsymbol{\alpha}_h \cdot \mathbf{y}_h$ is the highest order reflections, which input to the FDN *and* mix to the final output.

In figure 5.1 we see that the FDN does not take its input from the entire mixed output of the early reflections unit. Instead, it only takes input from the highest order of early reflections. The reasoning behind this is that the late reverb unit should continue the reflection model from where the early reflections unit left off. For example, if the early reflections are modeled up to second order, then the first set of impulses issuing from the FDN output represent the third order reflections, and so on.

We define the vector $\mathbf{y} = (y_1, y_2, ..., y_k)$ to be the list of outputs from the multi-tap delay shown in figure 5.2 at a given point in time. Additionally, we let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k)$ denote the list of gain coefficients that scale the output from the multi-tap delay.

In figures 5.1 and 5.2, only the highest order of early reflections provide input to the FDN. To distinguish the delay output taps and gain coefficients for the highest order reflections from those of lower order, we define some additional notation. Let $\mathbf{y}_l = (y_1, y_2, ..., y_l)$ be the first part of the vector $\mathbf{y}$ that represents the lower order reflections and let $\mathbf{y}_h = (y_{l+1}, ..., y_k)$ represent the vector of outputs representing the highest order reflections. In this way, the concatenation of $\mathbf{y}_l$ with $\mathbf{y}_h$ is the complete vector $\mathbf{y}$.

Similarly, we divide the gain coefficients vector $\boldsymbol{\alpha}$ into $\boldsymbol{\alpha}_l$, for the lower order reflections and $\boldsymbol{\alpha}_h$ for the higher order reflections.

The gain coefficients in the vector $\boldsymbol{\alpha}$ and the delay times of the output taps $\mathbf{y}$ are typically calculated by either the image source method or the Acoustic Rendering Equation (ARE). Any method of calculation that uses units of measurement consistently is acceptable. In this paper, we use the ARE (Siltanen et al., 2007).

The key problem with the structure shown in figure 5.1, with respect to modeling C_{80}, D_{50}, and T_S , is that the output taps coefficients α that model the highest order

early reflections represent sound rays as heard at the listener location. However, in an accurate reflections model, acoustic rays reflect off of the walls of the room, not the listener. In other words, by connecting the late and early reverb units together in this way, we are effectively placing the sound collection function (i.e. calculating acoustic intensity at listener location L due to reflections at infinitesimal surface area, see section 5.4.6 for definition of sound collection function) at listener position *before* the late reverb unit. In section 5.5 we show how this results in unrealistic energy output for the late reverb.

Our key contribution is the modified structure shown in figure 5.2 that takes two copies of the highest order early reflections $\mathbf{y}_h$ from the multi-tap delay in order to apply two different sets of gain coefficients to them. One copy mixes to the audio output after applying the gain coefficients in the vector $\boldsymbol{\alpha}_h$, and the other copy mixes using the gain coefficient vector $\boldsymbol{\beta}_h$ and becomes the input to the late reverb unit. The proposed structure is shown in figure 5.2. This change requires only one additional vector dot product to apply the second set of gain coefficients, $\boldsymbol{\beta}_h$, shown in figure 5.2 as the upper mixdown unit, and one multiplication with υ at the output of the FDN. Note that this multiplication by υ, which scales the overall output gain of the FDN could be avoided by grouping terms with the vector $\boldsymbol{\beta}_h$ but we show it placed after the FDN to emphasize that υ models the attenuation that happens after the last reflection as the energy is collected at the listener position. The key contribution is to efficiently compute the gain coefficient for each of those output taps. We discuss that computation in section 5.4.

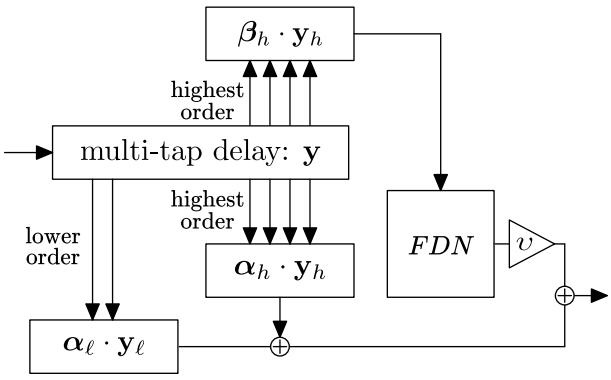

FIGURE 5.2: Block diagram of the proposed method. Unlike the method in figure 5.1, the delay taps representing highest order reflections $\boldsymbol{y}_h$ branch and multiply by two different scaling vectors, $\boldsymbol{\alpha}_h$ scales the signal that mixes to the final audio output and $\boldsymbol{\beta}_h$ scales the signal for the late reverb input. Having two different scaling vectors is important for sending the correct amount of energy into the late reverb unit because late reverb input is generally not equal to the early reflections output. Note that in practice the lower and highest order outputs of the early reflections unit are computed as a single dot product $\boldsymbol{\alpha} \cdot \boldsymbol{y}$ but we show it here with the output vector split into two sections y_ℓ and y_h to emphasize that the early part of the output is not sent to the FDN input.

5.4 METHOD

This section explains how to compute two energy-balancing coefficients in the proposed method: the vector $\boldsymbol{\beta}_h$, which scales the output of the multi-tap delay before it inputs to the late reverb unit and the scalar υ, which scales the output of the late reverb. This is illustrated in figure 5.2.

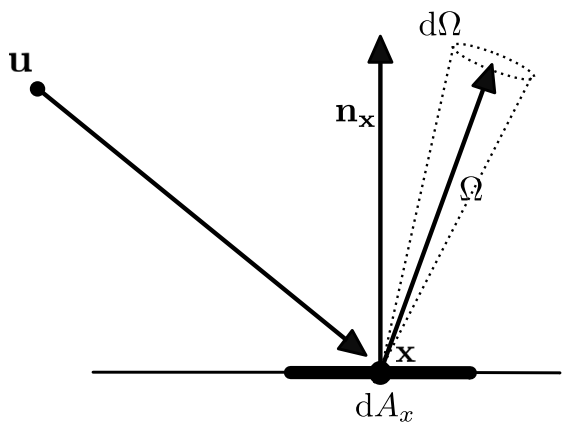

FIGURE 5.3: Radiance from the point $\boldsymbol{u}$ propagates toward the point $\boldsymbol{x}$, located on a differential unit of surface area, $\mathrm{d}A$. The acoustic radiance, $\ell(\boldsymbol{x}, \Omega) = \mathrm{d}\Phi/(\mathrm{d}\Omega\ \mathrm{d}A')$, quantifies the energy flux Φ reflected off $\boldsymbol{x}$ in the direction Ω per unit solid angle (steradian), per unit projected area A'. Note the following relation between area, A, and projected area, $A' = (\Omega \cdot \boldsymbol{n}_x)A$. The unit vector $\boldsymbol{n}_x$ is the surface normal.

5.4.1 The Acoustic Rendering Equation

We use the ARE (Siltanen et al., 2007) to model early reflections and calculate the scaling vectors $\boldsymbol{\alpha}_\ell$ and $\boldsymbol{\alpha}_h$, shown in figure 5.2. The key innovation with respect to our use of the ARE is the interpretation of the sound source in equation (5.8). Aside from that, our implementation uses a straightforward application of the ARE to model early reflections.

The ARE models the acoustic radiance, denoted by $\ell(\boldsymbol{x}, \Omega)$, at a point $\boldsymbol{x}$ emitted in the direction Ω. Radiance is a measure of outgoing energy flux per unit solid angle, per unit *projected* area (McCluney, 2014; Marschner, 2012; Nicodemus et al., 1977). Because the ARE borrows terminology from radiometry and computer graphics literature, and each field of study has its own idiosyncrasies of notation, there is some ambiguity among definitions and terms used in the related literature. To clarify them, the authors derived the relevant definitions from fundamental acoustic quantities and reviewed the related literature in (**AgusEnergyBased**).

In the paper that introduced the ARE, Siltanen et al. define the ARE as follows,

$$\ell(\boldsymbol{x}, \Omega) = \tag{5.1}$$

$$\ell_0(\boldsymbol{x}, \Omega) + \int_{\mathcal{G}} R(\boldsymbol{u}, \boldsymbol{x}, \Omega)\, \ell\left(\boldsymbol{u}, \frac{\boldsymbol{x} - \boldsymbol{u}}{|\boldsymbol{x} - \boldsymbol{u}|}\right) \mathrm{d}\boldsymbol{u}. \tag{5.2}$$

In the equation above, ℓ_0 is the emitted radiance and the integral term represents reflected radiance. Fig. 5.3 illustrates the physical meaning of the variables inside the integral. The integration region, $\mathcal{G}$, is the set of all points $\boldsymbol{u}$ in the surface geometry of the room and $\mathrm{d}\boldsymbol{u}$ is a differential unit of surface area. The function $R(\boldsymbol{u}, \boldsymbol{x}, \Omega)$ in the integral above is called the reflection kernel, defined in (Siltanen et al., 2007). It determines how much of the energy flux coming from the point $\boldsymbol{u}$ reflects off $\boldsymbol{x}$ in the direction Ω. The two simplest reflection kernels are pure specular, where all of the incoming energy flux from $\boldsymbol{u}$ reflects out at just one angle, and pure diffuse, where the energy flux from $\boldsymbol{u}$ spreads out evenly at all angles of reflection. When the reflection kernel is pure specular, the ARE is equivalent to the image source method (Siltanen et al., 2007). The product of the functions R and ℓ in the integral term of the ARE represents the component of the reflected energy flux at $\boldsymbol{x}$ going out in the direction Ω that derives its energy from an incident energy flux originating at point $\boldsymbol{u}$ elsewhere in the surface geometry of the room.

In our implementation, we discretise the surface geometry $\mathcal{G}$ into a set of discrete patches and use monte-carlo integration to compute the integral in equation (5.1) for each patch.

To simplify notation from here on, we define $\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}$ to be a unit vector pointing in the direction from $\boldsymbol{u}$ to $\boldsymbol{x}$,

$$\Lambda_{[\boldsymbol{u},\boldsymbol{x}]} = \frac{\boldsymbol{x} - \boldsymbol{u}}{\|\boldsymbol{x} - \boldsymbol{u}\|}. \tag{5.3}$$

Using this notation we rewrite equation (5.1) as follows,

$$\ell(\boldsymbol{x}, \Omega) = \tag{5.4}$$

$$\ell_0(\boldsymbol{x}, \Omega) + \int_{\mathcal{G}} R\left(\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}, \boldsymbol{x}, \Omega\right) \ell\left(\boldsymbol{u}, \Lambda_{[\boldsymbol{u},\boldsymbol{x}]}\right) \mathrm{d}\boldsymbol{u}. \tag{5.5}$$

5.4.2 Energy Flux Output of Point Sources

The definitions in this section are the first of three key ideas in this method. The goal of this section is to define point sources in a way that works consistently across two different acoustic modeling frameworks so that they can be combined in correct proportion to each other.

At the input of our acoustic model is a digital signal that represents a time series measurement of the sound pressure level in an acoustic medium, usually air. We begin with the assumption that the sound pressure is measured by a microphone positioned at a fixed distance d_M from a sound source, S. Although S moves freely within the model, the virtual microphone always remains at the same distance from S. This brings up a serious difficulty: the acoustic intensity of sound emitted at the point source S is infinite when measured at the point S itself. Therefore, we obviously can not allow

the sound source to have a distance of zero from other objects in the room. But if we require it to stay at least a minimum distance away from other objects, what should that minimum distance be? To answer this, we define the input signal p to be a measurement of the sound pressure level at a distance d_M from S, which we call the *minimum distance*, and we require that no object in the model can be placed closer to S than d_M. Physically, this is like saying that the microphone that recorded p was placed as close to the sound source S as possible, therefore no other object in the room can be closer than that. This ensures that no object in the room will observe a sound pressure level greater than $|p|$ coming from S.

This concept of minimum distance to the sound source implies a relationship between the audio input signal and the energy flux output of a sound source. Specifically, the intensity of acoustic energy flux, or Acoustic Intensity, of an isotropic sound source S, measured at a point $\boldsymbol{x}$ is,

$$I_a(S, \boldsymbol{x}) = \frac{p^2 {d_M}^2}{\|S - \boldsymbol{x}\|^2}. \tag{5.6}$$

Acoustic Intensity is a vector quantity. However, in (5.6) we simplify the notation by not writing the unit vector indicating its direction of propagation. Throughout this paper we define functions of energy flux without writing the direction vector because the direction is already evident from the function input arguments S and $\boldsymbol{x}$ and because subsequent calculations will use them as scalar quantities. Writing the definition in this way permits us to omit the vector magnitude symbol from the equations.

Radiant intensity is the measure of energy flux per unit solid angle (Marschner, 2012). At any point $\boldsymbol{x}$ located on the surface of the unit sphere around the isotropic sound source S, the acoustic intensity directed from S to $\boldsymbol{x}$, denoted by $I_a(S, \boldsymbol{x})$, is equal to the radiant intensity, denoted by $I_r(S, \boldsymbol{x})$,

$$I_r(S, \boldsymbol{x}) = I_a(S, \boldsymbol{x}) = p^2 {d_M}^2. \tag{5.7}$$

As in (5.6), we do not write the direction vector in the definition because it is evident from the function input arguments S and $\boldsymbol{x}$. Integrating the radiant intensity in (5.7) over a sphere solid angle, we find the total energy flux emitted by S as a function of the audio input, p,

$$\Phi(S) = \int_{4\pi} I_r(S, \boldsymbol{x})\, \mathrm{d}\Omega = 4\pi p^2 {d_M}^2. \tag{5.8}$$

The symbol 4π in the integration bounds above indicates integration over a sphere solid angle.

In the case where we want to define an anisotropic source, equation (5.7) can be modified into,

$$I_r(S, \boldsymbol{x}) = p^2 {d_M}^2 \zeta(\Omega), \tag{5.9}$$

where $\zeta(\Omega)$ will scale $I_r(S, \boldsymbol{x})$ according to direction Ω, such that equation (5.8) is still satisfied. Without loss of generality, we will use isotropic sound sources for the remainder of our explanation of this method.

5.4.3 Implications of the Minimum Distance

Equation (5.8) establishes a relationship between the minimum distance d_M and the total energy flux output of the sound source. This is key to understanding our method because our choice of d_M determines the overall volume of the reverb. If the listener and source are allowed to move freely around the room, then the loudest possible output volume occurs when the listener is at a distance of d_M from the source. Therefore, we must set the overall gain of the model to avoid clipping at that position. If we set a very small value for d_M, we need to set the output gain very low to avoid clipping, and we will get a lower volume output across all positions in the room. Therefore setting a low value for d_M is like saying that we recorded the audio input with a microphone very close to the source, so even when the volume is at its peak, it doesn't represent a very large total source energy flux. On the other hand, if we set a higher value of d_M, the total source energy flux is larger and we don't have to set the output gain so low to avoid clipping. However, we are forced to accept a more restricted range of motion for the source and listener.

5.4.4 Applying the ARE to Model Early Reflections

Our application of the ARE to model early reflections in this method is mostly standard. The unique feature is our use of the minimum distance d_M (defined in section 5.4.2) to quantify sound source energy flux in equation (5.10).

Emitted Radiance

By definition, the emitted radiance of the isotropic point source S in the direction Ω, denoted by $\ell_0(S, \Omega)$, is equivalent to the radiant intensity of S, denoted by $I_r(S, \Omega)$. Thus, we have the following expression for emitted radiance,

$$\ell_0(S, \Omega) = I_r(S, \Omega) = p^2 d_M^2. \tag{5.10}$$

Reflected Radiance

For first order reflections from a single point source S, the ARE as given in equation (5.4) reduces to the following,

$$\ell_1\left(\boldsymbol{x}, \Omega\right) = R(\Lambda_{[S,\boldsymbol{x}]}, \boldsymbol{x}, \Omega)\, \ell_0(S, \Lambda_{[S,\boldsymbol{x}]}), \tag{5.11}$$

which is simply the product of the reflection kernel and the emitted radiance of the point source.

Higher order reflections up to the Nth order are calculated recursively,

$$\ell_n(\boldsymbol{x}, \Omega) = \int_{\mathcal{G}} R(\Lambda_{[\boldsymbol{u},\boldsymbol{x}]}, \boldsymbol{x}, \Omega)\ell_{n-1}(\boldsymbol{u}, \Lambda_{[\boldsymbol{u},\boldsymbol{x}]})\mathrm{d}\boldsymbol{u}, \tag{5.12}$$

for all $n \in 1, 2, ..., N$. Because the expression above is recursive, for third order reflections and above we can speed it up by memoizing the lower order results. We calculate the integrals by monte-carlo method.

5.4.5 Late Reverb Energy Flux Input

In this section we explain how we calculate the gain coefficients that scale the amplitude of the multi-tap delay outputs that represent highest order reflections in order to provide the correct energy input to the late reverb unit.

The basis for this calculation is the assumption that energy decays exponentially and reflects diffusely. Although this assumption is not true for individual early reflections, when applied to the set of all early reflections as a group, it provides a useful estimate of the total acoustic energy flux in a room at a given point in time.

Initially, we experimented with the assumption that the energy level of each delay tap should scale according to a simple exponential decay envelope that depends on the timing of the reflection and nothing else. The results were unsatisfactory. The problem we encountered was that as the sound source moves around the room, it sends unequal portions of its energy flux output to the different surface patches in our model, depending on the distance between source and patch. Therefore, we need to scale the gain coefficients for the multi-tap delay output to represent not only the exponential decay envelope but also the proportion of the source energy flux reflecting off of each patch. This results in a method that resembles a simplified version of the ARE without the point collection function and the bidirectional reflectance distribution function ($BRDF$). We refer to the term $BRDF$ as defined in (Siltanen et al., 2007), of which the authors stated that the original term is borrowed from the field of optics (Nicodemus et al., 1977).

The ultimate goal of this part of the method is to compute the scaling vector $\boldsymbol{\beta}_h$, shown in figure 5.2. In the lower part of that figure we use the vector $\boldsymbol{\alpha}_h$ to scale the multi-tap delay outputs $\mathbf{y}$ for mixdown to the audio output. This quantity represents the highest order early reflections as heard from the position of the listener. In the upper part we take those same delay outputs, and scale them differently, so that $\boldsymbol{\beta}_h \cdot \mathbf{y}_h$ represents the highest order early reflections *from the position of the surface geometry*. The difference, therefore, between the two scaling vectors $\boldsymbol{\alpha}_h$ and $\boldsymbol{\beta}_h$ is that $\boldsymbol{\beta}_h$ doesn't include the point collection function (see Section 5.4.6) and the $BRDF$ in its calculation. Of course, reusing the delay outputs in this way means that the timing of individual reflections is not accurate. However, since the reverb unit doesn't model individual reflections, there is no need to accurately model individual reflection times at its input.

Energy Flux at the First Reflection

The goal of this section and section 5.4.5 is to model the total energy flux incident on the room surface geometry at the end of early reflections. We reason that the surface geometry energy flux should be the input to the late reverb unit because late reverb reflections come from the room surface geometry, not from the listener position, which is what would happen if we simply took the early reflections output ($\boldsymbol{\alpha}_h \cdot \boldsymbol{y}_h$) to be the late reverb unit input.

The method we use here borrows notation from the ARE. In our results section we show that we can apply the method presented here to an existing reverberator with only a five percent increase in computational time if we reuse some components of the early reverb calculation.

If we consider first order reflections, the proportion of energy flux that the source sends to each patch is determined by the size of the solid angle subtended by that patch as seen from the sound source. Likewise, the second and higher order reflected energy flux is determined by the size of the solid angle subtended by patch Y as seen from patch X.

So in summary, our method for estimating the energy flux input to late reverb is to assume that energy flux decays exponentially and reflects diffusely, at each reflection point dividing itself among the other surface patches in the model according to the size of the solid angle each patch subtends.

We use Sabine's formula (Schroeder et al., 2007; Kuttruff, 2009) to estimate the RT_{60} decay time for the room, given the room volume and absorption coefficients. Based on that decay time, we use the formula below, from Jot (Jot and Chaigne, 1991) to calculate the exponential reduction in amplitude of a signal as a function of the propagation distance between a pair of points $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\tau(\boldsymbol{x}, \boldsymbol{y}) = 10^{-3\,\|\boldsymbol{x}-\boldsymbol{y}\|\,/\,(c\,RT_{60})}, \tag{5.13}$$

where c is the speed of sound.

The total energy flux from S directed towards the surface patch X, denoted by $\Phi(S, X)$ is given by the following expression, similar to equation (5.8),

$$\Phi(S, X) = \int_X I_r(S, \Omega)\, \mathrm{d}\Omega, \tag{5.14}$$

where the differential unit $\mathrm{d}\Omega$ is a solid angle directed from S to X. The integration region X indicates the set of all solid angles Ω that start from S and point towards the surface X.

The same quantity can be expressed by area integration. To convert from integration by solid angle to integration by area, we define the function $P(S, \boldsymbol{x})$ whose numerator is the dot product of the direction of propagation and surface normal at $\boldsymbol{x}$ and whose denominator is the square of the distance,

$$P(S, \boldsymbol{x}) = \frac{\Lambda_{[S,\boldsymbol{x}]} \cdot \mathbf{n_x}}{\|S - \boldsymbol{x}\|^2}. \tag{5.15}$$

Recall that $\Lambda[S, \boldsymbol{x}]$ is a unit vector from S to $\boldsymbol{x}$. With this function we rewrite (5.14) as an area integral,

$$\Phi(S, X) = \int_X P(S, \boldsymbol{x})\, I_r(S, \boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{5.16}$$

$$= p^2 d_M^2 \int_X P(S, \boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{5.17}$$

In 5.16 each point on the surface X has a different distance from S and therefore a different propagation time and a different amount of exponential decay. To model the effect of this, we insert the term $\tau(S, \boldsymbol{x})$ that represents decay during propagation from S to $\boldsymbol{x}$. We use the notation $\phi(S, X)$ to denote the net energy flux incident on X from

S, accounting for dissipation and decay,

$$\phi(S, X) = p^2 d_M^2 \int_X \tau(S, \boldsymbol{x})^2 \; P(S, \boldsymbol{x}) \; \mathrm{d}\boldsymbol{x}. \tag{5.18}$$

For reflections beyond first order, we need to express the energy flux exchanged between patches. The function $g(\boldsymbol{x}, \boldsymbol{y})$ below is a geometry term that models visibility $\mathcal{V}$, distance $\|\boldsymbol{y} - \boldsymbol{x}\|^2$, and angle between surface normals,

$$g(\mathbf{x}, \mathbf{y}) = \mathcal{V}(\boldsymbol{x}, \boldsymbol{y}) \frac{(\Lambda_{[\mathbf{x},\mathbf{y}]} \cdot \mathbf{n}_\mathbf{x})(\Lambda_{[\mathbf{y},\mathbf{x}]} \cdot \mathbf{n}_\mathbf{y})}{\|\mathbf{x} - \mathbf{y}\|^2}. \tag{5.19}$$

Second and higher order reflections

For second order reflections, we have the following approximation of the energy flux at the second patch B that comes reflected off of the first patch A,

$$\phi(S, A, B) = \tag{5.20}$$

$$\frac{1}{\pi} \int_B \int_A \tau(\boldsymbol{a}, \boldsymbol{b})^2 \; g(\boldsymbol{a}, \boldsymbol{b}) \; \phi(S, \boldsymbol{a}) \; \mathrm{d}\boldsymbol{a} \; \mathrm{d}\boldsymbol{b}, \tag{5.21}$$

where $\boldsymbol{b}$ and $\boldsymbol{a}$ are points on the surfaces B and A. The expression under the integral is the product of the exponential decay τ, the geometry term g, and the first order energy $\phi(S, \boldsymbol{a})$.

For third order reflections the energy flux is,

$$\phi(S, A, B, C) = \tag{5.22}$$

$$\frac{1}{\pi} \int_C \int_B \tau(\boldsymbol{b}, \boldsymbol{c})^2 \; g(\boldsymbol{b}, \boldsymbol{c}) \; \phi(S, A, \boldsymbol{b}) \; \mathrm{d}\boldsymbol{b} \; \mathrm{d}\boldsymbol{c}. \tag{5.23}$$

And similarly, for fourth order it is,

$$\phi(S, A, B, C, D) = \tag{5.24}$$

$$\frac{1}{\pi} \int_D \int_C \tau(\boldsymbol{c}, \boldsymbol{d})^2 \; g(\boldsymbol{c}, \boldsymbol{d}) \; \phi(S, A, B, \boldsymbol{c}) \; \mathrm{d}\boldsymbol{c} \; \mathrm{d}\boldsymbol{d}. \tag{5.25}$$

The pattern continues in the same way for higher orders of reflection.

We further simplify the notation by using a list to represent the input arguments of the function ϕ. For example, instead of writing $\phi(S, A, B, C)$ we write $\phi[F]$, where $F = [S, A, B, C]$. For each multi-tap delay output $\boldsymbol{y}_i$ in the early reflections module there is a specific list of surface geometry reflections that indicates the path that reflection took to move through the model. We use the symbol F_i to represent the list of reflection points that corresponds to the reflection modeled by the delay output $y_i \in \boldsymbol{y}$.

Since ϕ is defined recursively for higher order reflections, then it always contains the factor of p^2 that appears in the base case of the recursion, equation (5.18). This is significant because p^2 changes with each input sample and therefore it must be factored out of the energy flux equations for implementation of the method. We define the function $\phi'(F_i)$ to be the ratio of the energy flux $\phi(F_i)$ where the ith reflection meets the

wall, to the energy flux p^2 at the microphone location,

$$\phi'(F_i) = \phi(F_i)/p^2. \tag{5.26}$$

This removes the dependence on p from the equation so that ϕ' is a constant for any stationary listener and source position. Finally, the elements $\beta_{h:i}$ of the scaling vector $\boldsymbol{\beta}_h$ are given by the following expression,

$$\beta_{h:i} = \sqrt{\int_X \tau(\boldsymbol{x}, L)\, \phi'(F_i)\, \mathrm{d}\boldsymbol{x}}, \tag{5.27}$$

where X is the last surface patch in F_i. The term $\tau(\boldsymbol{x}, L)$ needs more detailed explanation. It might look like this is the energy decay for a signal going from a point $\boldsymbol{x}$ on the wall to the listener position at L, but that is not the intention. Remember that the goal of calculating $\beta_{h:i}$ is to estimate the energy of the ith reflection as it intersects the room surface geometry, not the listener. However, the timing of the output taps in the early reflections unit corresponds to reflections that reach the listener, so if we stop the model when we reach the last surface patch in the list F_i then we have not followed our initial assumption that the energy decays exponentially over time. We multiply the $\tau(\boldsymbol{x}, L)$ term here to account for the exponential decay that corresponds to the additional propagation time from the last reflection listed in F until the early reflections unit actually outputs $\boldsymbol{y}_i$.

Note that this $\tau(\boldsymbol{x}, L)$ is not a collection function at L and that we do not use any collection function in the calculation of $\beta_{h:i}$.

5.4.6 Late Reverb Energy Output

In this section we introduce how to calculate the late reverb output scaling coefficient, υ, shown at the output of the FDN in figure 5.2. The purpose of υ is to represent the ratio between total energy flux in the room and acoustic intensity at the listener location. So it represents a point collection function and a change of units at the same time.

Borrowing from the notation of the ARE, we define $\ell_+(\boldsymbol{x}, \Omega)$ to be the total late reverb radiance from surface $\boldsymbol{x}$ directed parallel to the unit vector Ω.

Our late reverb reflections must be energy preserving because energy losses are already modeled by the FDN (**AgusEnergyBased**). This implies that the total irradiance at a point $\boldsymbol{x}$ is equal to total radiance. In other words, energy input equals energy output. Therefore the following is our conservation of energy requirement at all surface points $\boldsymbol{x}$,

$$E_+(\boldsymbol{x}) = \int_{2\pi} (\boldsymbol{n}_{\boldsymbol{x}} \cdot \Omega)\, \ell_+(\boldsymbol{x}, \Omega)\, \mathrm{d}\Omega, \tag{5.28}$$

where $E_+(\boldsymbol{x})$ is the late reverb irradiance at $\boldsymbol{x}$. Irradiance is incoming energy flux per unit area. Acoustic irradiance is defined in (Siltanen et al., 2007) and the original concept of irradiance in the context of radiometry is defined in (Nicodemus et al., 1977).

We assume diffuse reflection for late reverb. This allows us to take $\ell_+(\boldsymbol{x}, \Omega)$ out of the integral and simplify,

$$E_+(\boldsymbol{x}) = \ell_+(\boldsymbol{x}, \Omega) \int_{2\pi} \boldsymbol{n_x} \cdot \Omega \, \mathrm{d}\Omega, \tag{5.29}$$

$$= \ell_+(\boldsymbol{x}, \Omega) \, \pi. \tag{5.30}$$

Rearranging terms, we have the following approximation for the late reverb radiance at $\boldsymbol{x}$,

$$\ell_+(\boldsymbol{x}, \Omega) = \frac{1}{\pi} E_+(\boldsymbol{x}), \tag{5.31}$$

which amounts to saying that energy reflects diffusely at $\boldsymbol{x}$ and incoming and outgoing energy flux are equal.

Because the FDN mixes energy evenly to all its delay lines, we work with the assumption that late reverb energy is evenly mixed in the room. This implies that $E_+(\boldsymbol{x}) = E_+$, in other words E_+ doesn't depend on $\boldsymbol{x}$. This is related to the idea that the average intensity of late reverberation reaching the listener over a finite time interval is approximately the same in every direction (Angel, Algazi, and Duda, 2002). Griesinger's research supports this finding by stating that the late reverberation is so well mixed that the average amplitude of reverberation at every point along the wall is approximately the same (Griesinger, 2000).

Here we define the point-collection-function, $h(\boldsymbol{x}, L)$ to represent the acoustic intensity at the listener location L due to reflections at the surface geometry point $\boldsymbol{x}$. The purpose of the point collection is to convert from units of radiance at $\boldsymbol{x}$ to units of incident energy flux per unit area at L. The function $h(\boldsymbol{x}, L)$ has two components: a visibility term, $\mathcal{V}(\boldsymbol{x}, L)$, and a point-listener geometry term, $P(\boldsymbol{x}, L)$,

$$h(\boldsymbol{x}, L) = \mathcal{V}(\boldsymbol{x}, L) \, P(\boldsymbol{x}, L). \tag{5.32}$$

The binary visibility function, $\mathcal{V}(\boldsymbol{x}, L)$, is one if L is visible from $\boldsymbol{x}$ and zero otherwise. The geometry term P in the point collection function was defined in equation (5.15).

The following expression expresses the irradiance at L reflected from the entire room surface geometry, $\mathcal{G}$,

$$E_+(L) = \int_{\mathcal{G}} h(\boldsymbol{x}, L) \, \ell_+(\boldsymbol{x}, \Omega) \, \mathrm{d}\boldsymbol{x}. \tag{5.33}$$

Let $\Phi(FDN)$ be the total energy output of the FDN. Since the energy output of the FDN represents the total energy reflected at all surfaces then the average reflected energy flux per unit area is the total FDN output flux divided by total surface area,

$$\frac{\Phi(FDN)}{\mathcal{G}} = \text{late reverb energy output per unit area.} \tag{5.34}$$

Combining this with equation (5.31), we have the following expression for irradiance at the listener position,

$$E_+(L) = \frac{\Phi(FDN)}{\pi \mathcal{G}} \int_{\mathcal{G}} h(\boldsymbol{x}, L)\, \mathrm{d}\boldsymbol{x}. \tag{5.35}$$

We define υ^2, to be the ratio of irradiance at the listener to energy flux output of the FDN,

$$\upsilon^2 = \frac{E_+(L)}{\Phi(FDN)} \tag{5.36}$$

Combining (5.36) with (5.35), we have the following final expression for υ,

$$\upsilon = \sqrt{\frac{1}{\pi \mathcal{G}} \int_{\mathcal{G}} h(\boldsymbol{x}, L)\, \mathrm{d}\boldsymbol{x}}. \tag{5.37}$$

5.4.7 Method Summary

In overview, the proposed method has the three steps listed below. In each step we highlight the key contributions of our method with italic text.

1. Obtain the vector of gain scaling coefficients $\boldsymbol{\alpha}$ using the ARE. This vector scales the multi-tap delay outputs for mixing to the final audio output. This part of our method uses the ARE to model early reflections in the usual way, except for one difference: *our definition of energy flux from a point source in equation (5.8) ensures consistency of units when we use the ARE together with other modeling methods*.

2. *Calculate $\boldsymbol{\beta}_h$ using equation (5.27).* The vector $\boldsymbol{\beta}_h$ scales the each of the multi-tap delay outputs to provide the late reverb unit with a correctly modeled energy flux input. Existing methods use the same scaling vector $\boldsymbol{\alpha}$ for both final audio output and late reverb input. However, in an accurate physical model, the output of early reflections and input of late reflections generally do not have the same energy flux, so two different scaling vectors $\boldsymbol{\alpha}_h$ and $\boldsymbol{\beta}_h$, as shown in figure 5.2, are necessary.

3. *Calculate υ using equation (5.37).* The scalar variable υ scales the late reverb output, representing the collective effect of the point collection function that models collection of late reverb energy at the listener position, as it applies to the entire set of all late reflections propagating from the room surface geometry to the listener.

	R1 P1	R1 P2	R1 P3	R2 P2	R2 P3	R2 P4	R3 P1	R3 P2	R4	R5	σ	JND	$\frac{\sigma}{1JND}$
D_{50}	0.95	0.93	0.92	0.63	0.50	0.36	0.89	0.74	0.46	0.51	0.22	0.05	4.48
C_{80} (dB)	18.48	16.51	15.11	5.77	4.23	2.59	12.09	9.03	0.55	1.03	6.67	1	6.67
T_S (ms)	16.05	17.90	22.30	55.95	70.35	85.15	21.40	37.35	108.6	151.0	45.36	10	4.54
RT_{60} (s)	0.34	0.36	0.36	0.85	0.87	0.90	0.56	0.56	1.78	3.24	0.90	Rel. 5%	18.4
EDT (s)	0.18	0.22	0.22	0.77	0.81	0.87	0.37	0.52	1.73	3.50	1.02	Rel. 5%	22.18

TABLE 5.2: D_{50}, C_{80}, T_S, RT_{60} and EDT values of all 10 chosen RIRs, averaged in the 500Hz and 1000Hz octave bands.

5.5 EVALUATION

5.5.1 RIR Recording and Simulation

We use room impulse responses (RIRs) from five different rooms, three of which are taken from the AIR database (Jeub, Schäfe, and Var, 2009):

1. R1 (AIR): A meeting room (8.0m length, 5.0m width, 3.10m height, $124m^3$) with glass windows in the room and walls that are made of concrete. The room has an average RT_{60} time of 0.23s.

2. R2 (AIR): A lecture room (10.8m width, 10.9m length, 3.15m height, $412m^3$) used for seminars with common equipments such as desks and chairs. The wall surface is consisted of 3 glass windows and 1 concrete wall. Its average RT_{60} time is 0.78s.

3. R3 (AIR): A small office room (5.0m width, 6.4m length, 2.9m height, $92.8m^3$) with common office equipments in the vicinity. The room has an average RT_{60} time of 0.43s.

4. R4: An empty, almost rectangular room (1.89m width, 5.58m length, 3m height, $31.6m^3$) with concrete ceiling and floors, and polished marble walls. The room is a basement lift lobby. There are three alcoves in the walls, which serves as a lift opening. The lift doors were closed throughout the recording of the impulse response. The average RT_{60} is 1.78s.

5. R5: An empty, almost rectangular room with concrete walls, ceiling, and floor (10m width, 16m length, 4.3m height, $688m^3$). The room has several glass windows, and an average RT_{60} time of 3.2s.

We generated the RIRs in R4 and R5 using the logarithmic sine sweep method (Farina, 2000), with a 50 second sweep between 50 and 20000 Hz. We recorded those RIRs using an omni-directional microphone. After deconvolution, the resulting RIRs are truncated using Lundeby's method (Lundeby et al., 1995) in order to prevent noise from affecting our measurements of decay time. All RIRs exceed the minimum decay

range of 57 dB, as recommended in (Hak, Wenmaekers, and Luxemburg, 2012). In order to calculate room acoustic parameters presented in section 5.5.2, we find the cross-point of the decay time, which is the time when the impulse response crosses below the noise floor using the Lundeby's method (Lundeby et al., 1995; Hak, Wenmaekers, and Luxemburg, 2012).

To obtain the simulated RIRs, we implemented our proposed method in C++ in an iOS application that can process both pre-recorded and live audio input in real time. We tested the software on an iPad Air 2 simulator (running on a Mac laptop with 2.5 GHz Intel Core i7 CPU and 16GB RAM). We simulated each of the recorded RIRs a virtual room subdivided into 24 discreet patches, modeling early reflections up to the second order. To compute the integrals in the method section of this paper, we used Monte Carlo simulation with 100 points per patch. In total, we simulated 10 different RIRs, each from a different room configuration: several source-microphone positions in R1 (P1 to P3), R2 (P2 to P4), and R3 (P1 and P2), and one position in each R4 and R5. The geometric locations of sources and microphone in R1 corresponds to source-microphone distance of 1.45m (P1), 1.7m (P2), and 2.8m (P3) in AIR database (Jeub, Schäfe, and Var, 2009). In R2, P2 to P4 corresponds to source-microphone distance of 4m, 5.56m, and 7.1m respectively and in R3, the source-microphone distance we picked is 1m (P1) and 2m (P2) (Jeub, Schäfe, and Var, 2009). In both R4 and R5, both the source and microphone is placed in the middle of the room, with a distance of 0.7m and 1.5m between them respectively. Throughout the rest of this section we analyze those ten simulated RIRs to evaluate our method.

We do not present evaluation results using RIRs from Aula Carolina mentioned in section 5.2.1, as its shape is much more complex and requires advanced geometric modeling, of which such data is not publicly available. R5 serves a substitute for a bigger rooms with long reverberation time. For a more compact representation of the data in this section, we picked a representative subset of configurations (in terms of all room parameters) in R1 (meeting room), R2 (lecture room), and R3 (office room), instead of using all of their publicly available IRs with source-microphone configurations presented earlier in table 5.1. Table 5.2 summarizes all common room parameter values (Iso3382-1, 2009), along with their standard deviations and JNDs. The value of $\sigma/(1JND)$ is larger than 4 for all room parameters, which indicates a good variation in the selected IRs.

To see how much our proposed model improves the accuracy of the model, we simulated the same set of 10 impulse responses using the model shown in Figure 5.1 as a baseline for comparison. Except for the addition to the proposed method of the scaling vector $\boldsymbol{\beta}_h$ as shown in figure 5.2, the baseline and proposed method implementations are exactly the same. With respect to the early / late reverb energy balance, our baseline method closely resembles the structures used in (Jot, 1997; Menzer, 2010; Wendt, Par, and Ewert, 2014a; Wendt, Par, and Ewert, 2014b; Carty and Lazzarini, 2010; Sarti and Tubaro, 2001; Savioja et al., 1999), where the output of the highest order early reflections is connected directly into the late reverberation unit.

The FDN used for both baseline and proposed methods has 16 delay lines, using a Fast Hadamard Transform in place of the unitary feedback matrix. This amount of delay lines is sufficient according to (Jot, 1997). The average delay length in the FDN is slightly longer than the mean free path in the room (Smith, 2010).

In all, we have ten sets of three impulse responses. Each set comprises a recorded RIR, a RIR simulated by the proposed method and an RIR simulated using the baseline method. The rest of this section presents measurements on those RIRs of acoustic parameters defined in (Iso3382-1, 2009).

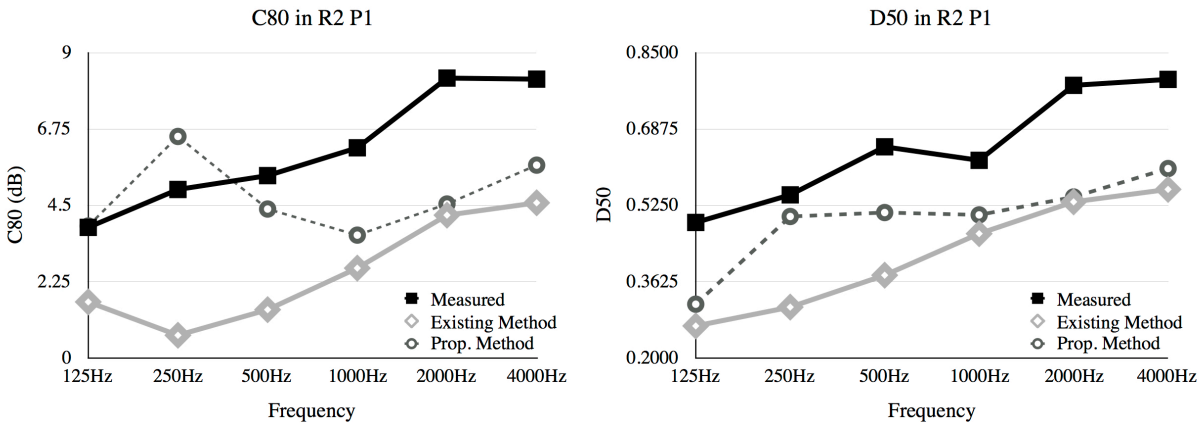

FIGURE 5.4: The plots of C_{80} and D_{50} in R2 P2 across 6 frequency bands from 125Hz to 4000Hz. Black line: measured RIR, gray line: baseline method, dashed line: proposed method.

5.5.2 Results

This section discusses how our proposed method compares to the baseline method, in terms of D_{50}, C_{80}, T_S. All room parameter values (D_{50}, C_{80}, T_S, RT_{60} and EDT) presented in this section are reported in terms of absolute value of JND with respect to the recorded RIR. As mentioned before, the JND for D_{50}, C_{80}, and T_S is 0.05, 1dB, and 1ms respectively (Iso3382-1, 2009). They are averaged in the 500Hz and 1000Hz octave bands and obtained from an arithmetic average of 15 iterations on each room configuration. We also report how much additional calculation time our method requires.

Decay Time

Decay time directly influences Clarity, Definition, and Centre Time, which are the main measures we use to evaluate our method. Typically we would use Sabine's formula to estimate the decay time of an unknown room. But in this evaluation, we are comparing our simulation against recorded RIRs, so we can exactly compute the RT_{60} decay time of each room by measuring it directly from the RIR. We then set the RT_{60} decay of the FDN in both the baseline simulation and proposed method simulation exactly equal to the decay time of the recorded RIR from each room we simulated. This ensured that errors in the estimate of the decay time did not distort our evaluation of Clarity, Definition, and Centre Time for the simulations.

Table 5.4 shows the raw values of EDT for measured RIR, baseline simulation, and proposed method simulation. EDT has a comparatively higher average JND values when compared to the other room parameters: D_{50}, C_{80}, and T_S. One possible reason is that EDT is known to be more sensitive towards any possible flaws and errors in modeling and numerical approximation (Iso3382-1, 2009). Although both existing and proposed methods do not explicitly model the early decay time, it is interesting to note that the average absolute JND of the proposed method is about 1.5 times more accurate than the existing method. Furthermore we can observe that the proposed method has comparatively lower JND than the existing method in all except one room configuration (R5).

Early to Late Reverb Energy Balance

In table 5.3, we compare our proposed method to the baseline method in terms of three room acoustic parameters that measure the balance between early and late energy: D_{50}, C_{80}, and T_S. The values presented in table 5.3 are an average of the 500 and 1000Hz frequency bands, which are the standard frequency bands for these measurements (Iso3382-1, 2009). On average, our proposed method is more than twice as accurate as the baseline method.

In table 5.3 we show the raw values of D_{50}, C_{80}, and T_S for measured RIR, baseline simulation and proposed method simulation. Simulation errors for both methods are expressed in units of JND and we show the ratio of error between the baseline method and proposed methods.

The standard deviation of the error is reported in units of JND. We observe a standard deviation of 1.37 (D_{50}), 0.94 (C_{80}), and 1.12 (T_S) for the baseline method, reduced to 0.81 (D_{50}), 0.58 (C_{80}), and 0.47 (T_S) for the proposed method. In all source-listener position configurations we tested, we observed no case where the proposed method performed worse than the baseline method.

The value of 10% trimmed mean of the factor of improvement is 2.29 (D_{50}), 2.79 (C_{80}), and 2.57 (T_S), which indicates that the simulations from the proposed method are still at least two times more accurate than the baseline method even after we exclude the cases where the improvement is maximum and minimum. Figure 5.4 plots the value of C_{80} and D_{50} in R2 P2 in octave bands from 125Hz to 4000Hz. It shows that the proposed method is generally able to improve the accuracy of these metrics across various frequency bands. For both C_{80} and D_{50} values, the proposed method (dashed line) is generally closer to the measured RIR (black line) throughout all frequency bands, as compared to the baseline method (gray line).

We performed a one-tailed t-test with an alternate hypothesis of $\log \left| \frac{\text{baseline JND}}{\text{proposed JND}} \right| > 0$. In other words, we tested the hypothesis that the proposed method is more accurate than the baseline. For all four parameters, C_{80}, D_{50}, T_S, and EDT we reject the null hypothesis with greater than 99 percent confidence. The p-values are 0.0012, 0.000023, 0.00001, and 0.0056 for D_{50}, C_{80}, T_S, and EDT respectively.

Table 5.3 expresses the effect of the proposed method in terms of ratios. In absolute terms, the values measured for that table tend to increase proportional to the decay time of the room. Therefore the improvement, in absolute terms, is typically greater for rooms with long decay time and less significant in rooms with short decay time. This

is significant for modeling applications where only short decay times are needed; in those cases, it may be recommended not to implement the proposed method because the improvement will likely be inaudible. On the other hand, for very long decay times, the improvement achieved by implementing the proposed method is more significant. Table 5.1 shows the average improvement achieved by the proposed method in four rooms. Sections (a) - (d) of table 5.1 are sorted with (a) having the longest decay time and (d) having the shortest decay. We can see there that the improvement is clearly audible for rooms (a) - (c) but in room (d) with the shortest decay time, whether or not the improvement would be audible is questionable.

Calculation Time

	S1	S2	S3
	Baseline Method		
Avg. time (s)	14.3	72.0	19.02
σ time (s)	0.22	0.72	0.35
	Proposed Method		
Avg. time (s)	15.1	75.9	20.0
σ time(s)	0.18	0.49	0.44

TABLE 5.5: The average time taken and its standard deviation to render impulse responses from all 10 configurations using baseline and proposed method.

To investigate how much additional running time is introduced by the proposed method, we run the simulations of each room configuration with three different level-of-detail settings:

1. S1: using 24 patches and doing numerical integration with 100 sample points one each patch
2. S2: using 54 patches and doing numerical integration with 100 sample points one each patch
3. S3: using 54 patches and doing numerical integration with 50 sample points one each patch

The average time required to render all 10 RIRs using the baseline and the proposed method are presented in table 5.5, along with their standard deviations. We measured the computation times by running C++ implementations of the baseline and proposed models in offline processing mode and measuring the time to produce 30 impulse responses for each of the three configurations S1, S2, and S3. The proposed method is sufficiently similar to the ARE (Siltanen et al., 2007) so that we can memoize and reuse data from some parts of the early reflections ARE calculation to speed up the late reverb energy input calculation.

The differences in calculation time in seconds between the baseline and proposed method are shown as percentages in table 5.6. On average, the proposed method requires less than 5.5% additional calculation time. The maximum additional time needed to run the proposed method is 7.64% (R4, S1). This indicates that the proposed method could be added on to existing real time methods without loosing the ability to operate in real time. Also, approximately the same amount of additional time is needed for all three settings.

It is important to note that the results in table 5.6 are measured in a context where the baseline method and the proposed method were both using the ARE to model early reflections. This is significant because the ARE requires many of the same numerical calculations as the method we use to calculate late reverb energy flux. We made use of the similarity to reduce the computation time by memoizing partial results of the ARE early reflections computation and reusing them for the late reverb energy flux estimate. If the baseline method had used the image source method, we would not have been able to reuse so much of the data from the early reflections to speed up the late reverb model. However in this case, the runtime of the method we propose is still in the same asymptotic complexity class as the early reflections; it would differ only by a constant multiple of the baseline performance time.

	S1 Δt(%)	S2 Δt(%)	S3 Δt(%)
R1 P1	5.11	5.19	5.65
R1 P2	5.40	5.89	4.14
R1 P3	3.96	6.16	3.46
R2 P2	5.82	6.36	5.40
R2 P2	2.64	4.00	6.18
R2 P4	7.52	5.36	5.11
R3	7.64	5.44	6.88
R4	4.60	5.77	4.67
R5	4.35	4.95	4.89
Average	**5.23**	**5.46**	**5.15**

TABLE 5.6: Percent increase in computation time for proposed method compared to baseline method. The symbols S1, S2, and S3 indicate three different levels of modeling detail.

5.6 CONCLUSION

In this paper we introduced a method to compute energy modeling coefficients that can be applied to existing hybrid acoustic modeling reverberators. The first set of coefficients, in the vector $\boldsymbol{\beta}_h$, are applied at each of the output taps of the highest order early reflections, while the second coefficient, υ, is applied at the single channel mixed-down output of the FDN. We used the exponential decay formula from (Jot and Chaigne, 1991) and the ARE (Siltanen et al., 2007) to compute $\boldsymbol{\beta}_h$ and υ.

In section 5.5 we showed that on average, the proposed method improves the accuracy of the baseline method by at least a factor of 2 in terms of D_{50}, C_{80}, and T_S, with

less than 5.5% increase in computation time. This means that it will most likely not hinder the ability of the baseline method to run in real time.

5.7 FUTURE WORK

In (Välimäki et al., 2012), Valimaki et. al mentioned that delay networks such as the FDN appear to be the most efficient yet perceptually convincing artificial reverberators, and therefore it is no surprise that it remains widely used by many of the hybrid reverberation algorithms to approximate late reflections. However there exist many other types of delay networks, such as the Scattering Delay Network proposed in (Sena et al., 2015). Convolution is also a widely used option for the type of late reverb unit used here. Hence, future work may include applying the idea of late reverb energy flux modeling to related methods that replace the FDN with other kinds of late reverb structures.

	Measured	Baseline Method	Baseline Method |JND|	Proposed Method	Proposed Method |JND|	Improveme $\left\lvert\frac{\text{Baseline JND}}{\text{Proposed JND}}\right\rvert$
				D_{50}		
R1 P1	0.947	0.928	0.39	0.950	0.06	6.91
R1 P2	0.934	0.902	0.64	0.928	0.11	5.61
R1 P3	0.918	0.841	1.54	0.870	0.95	1.61
R2 P2	0.634	0.466	3.37	0.558	1.52	2.22
R2 P3	0.501	0.311	3.80	0.389	2.24	1.7
R2 P4	0.361	0.257	2.08	0.292	1.38	1.51
R3 P1	0.887	0.767	3.20	0.799	2.58	1.24
R3 P2	0.744	0.678	2.40	0.697	1.76	1.37
R4	0.459	0.353	2.11	0.394	1.29	1.63
R5	0.511	0.377	2.67	0.461	0.99	1.24
Average			2.42		1.29	**2.65**
σ			1.37		0.81	
				C_{80}		
R1 P1	18.48	16.37	2.12	17.90	0.59	3.6
R1 P2	16.51	14.88	1.63	15.73	0.78	2.08
R1 P3	15.11	12.76	2.35	13.98	1.13	2.08
R2 P2	5.77	2.58	3.19	5.04	0.73	4.39
R2 P3	4.23	0.68	3.56	3.14	1.10	3.24
R2 P4	2.59	-1.22	3.82	1.48	1.12	3.41
R3 P1	12.09	9.02	3.07	9.66	2.43	1.26
R3 P2	9.03	7.26	1.77	7.81	1.21	1.46
R4	0.55	-0.31	0.85	0.28	0.27	3.15
R5	1.03	-1.21	2.24	0.35	0.68	3.28
Average			2.46		1.00	**2.8**
σ			0.94		0.58	
				T_S		
R1 P1	0.016	0.020	0.41	0.018	0.17	2.38
R1 P2	0.018	0.023	0.56	0.020	0.26	2.18
R1 P3	0.022	0.031	0.84	0.030	0.72	1.16
R2 P2	0.056	0.070	1.44	0.057	0.14	10.29
R2 P3	0.070	0.088	1.73	0.075	0.42	4.11
R2 P4	0.085	0.103	1.76	0.089	0.36	4.89
R3 P1	0.021	0.035	1.37	0.031	1.01	1.36
R3 P2	0.037	0.046	0.88	0.044	0.68	1.3
R4	0.109	0.125	1.69	0.119	1.02	1.65
R5	0.151	0.195	4.36	0.167	1.60	2.72
Average			1.50		0.64	**3.2**
σ			1.12		0.47	

TABLE 5.3: Raw measurements along with absolute value of error for definition (D_{50}), clarity (C_{80}), and centre time (T_S) and across all 10 room configurations for baseline and proposed methods, measured in units of JND. The symbol σ indicates the standard deviation. The proposed method is on average more than twice as accurate as the baseline method.

	Measured	Baseline Method	Baseline Method |JND|	Proposed Method	Proposed Method |JND|	Improvement: $\left\lvert \frac{\text{Baseline JND}}{\text{Proposed JND}} \right\rvert$
				EDT		
R1 P1	0.176	0.249	8.25	0.202	2.91	2.83
R1 P2	0.218	0.291	6.73	0.253	3.27	2.06
R1 P3	0.224	0.343	10.64	0.313	7.99	1.33
R2 P2	0.773	0.964	4.95	0.834	1.59	3.12
R2 P3	0.808	0.931	3.05	0.862	1.33	2.3
R2 P4	0.874	1.017	3.29	0.771	2.34	1.4
R3 P1	0.368	0.525	8.55	0.498	7.10	1.21
R3 P2	0.517	0.637	4.68	0.637	4.68	1.0
R4	1.725	1.767	0.49	1.753	0.32	1.51
R5	3.497	3.397	0.57	3.293	1.17	0.49
Average			5.12		3.27	**1.72**
σ			3.41		2.57	

TABLE 5.4: Raw measurements along with absolute value of error for early decay time (EDT) and across all 10 room configurations for baseline and proposed methods, measured in units of JND. The symbol σ indicates the standard deviation.

Chapter 6

Adaptive Lateral Room Patch Decomposition for Binaural Room Modeling

6.1 ABSTRACT

For our work in the previous chatpers, 4 and 5, we subdivide the surfaces in the room evenly for later computation using the ARE. In other words, each surface *patch* is of roughly the same size. In the case that there are numerous polygons in the existing 3D model, we grouped the polygons into equal group sizes before performing Monte Carlo computation to solve the ARE. However in (Kuttruff, 2009), it is stated that human hearing is most sensitive to reflections that come from lateral directions (left and right direction at the ear level) rather than from the front, or back. Furthermore, according to a study by Barron et. al (Barron and Marshall, 1981), the sense of spaciousness is only largely due to these lateral reflections. Therefore we theorize that we may simplify computations for the rest of the surface patches that do not contribute reflections directly in the lateral directions if we have more geometrical details on the surfaces that intersects the lateral plane of the listener at the ear level and. In this chapter we present a method that gives finer subdivision of surfaces near the listener ears and we plan to investigate whether this method further improves the accuracy and computational time of the work in section 5.

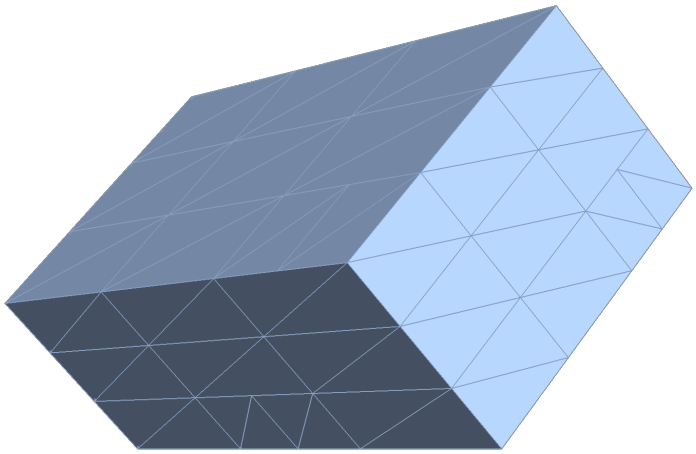

FIGURE 6.1: Approximately even subdivision on walls in rectangular room.

6.2 BACKGROUND AND MOTIVATION

One of the ways to solve the integral in the ARE is by performing Monte-Carlo simulation on each of the surface patches in the 3D model of the room. The size of each patch is typically similar, such as shown in (Bai, Richard, and Daudet, 2015; Raghuvanshi, Narain, and Lin, 2009). In (Raghuvanshi, Narain, and Lin, 2009), some surfaces are further divided if there is enough memory space. Our work in Chapter 4 also evenly divide room walls into patches, e.g: divide a rectangular room into 54 patches, 9 patches per wall. Since we require N to be power of two, if $N = 64$, we randomly select 10 more patches to be further divided into smaller patches, as shown in Figure 6.1.

However if we do not carefully subdivide the patches, some channels may have no output. Recall in Chapter 4 that the multiplexer groups the energy output from each patch into channel based on listener's azimuth. For example, one may evenly subdivide surfaces such as walls as shown in figure 6.2 and divide the listener's azimuth into 16 channels. Figure 6.2 shows the top-view of a listener placed in a rectangular room (not to scale). Each colored line represent a surface patch. We would set the midpoint of each patch (indicated as black circles) as the representative point of that patch (used when computing delay times of each sound rays). As a result, channel $0°$ to $22.5°$, $45°$ to $67.5°$, and many other channels (half of the total channels to be exact) do not have any output from lateral reflections. This scenario is more apparent in the case that the listener ear is placed near to a wall, such as less than 30 cm of distance.

Barron et. al (Barron and Marshall, 1981) proposes that the lateral reflections give the sense of spaciousness in the room. It is also stated in (Kuttruff, 2009) that humans are most sensitive to reflections coming from lateral directions, i.e: directly on the entrance of ear canal such as on the axis of external auditory canal. Brown et. al (Brown and Duda, 1998) designed pole-zero first-order HRTF filter which behavior can be set to boost the signal the most at azimuth $+100°$ or $-100°$, the typical entrance of the right and left ear canal. This shows that reflections from the lateral direction contain critical perceptual cues.

In the next section, we propose a new method to address this issue and also allow

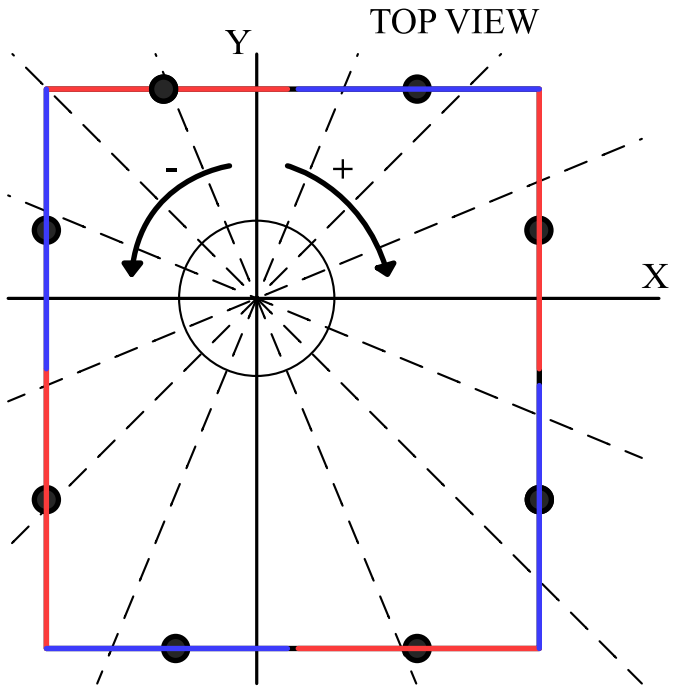

FIGURE 6.2: Issue that may rise from even wall subdivision. Some azimuthal section does not have an output, illustrated by the black dots.

the auralization to be more flexible since our method does not require the number of patches in the room to be equal to be the number of delay lines in the FDN (typically is a power of two if Hadamard matrix is used). We do this by grouping the surface patches or polygons and model each group with one FDN delay line.

6.3 METHOD

Studies by Barron et. al (Barron and Marshall, 1981) shows that the sense of spaciousness is proportional to $E \cos\theta$, where E is the energy of the sound waves and ϕ is the elevation angle in the typical vertical-polar coordinate system (see Figure 6.3. We propose the following method to do ensure that when a listener is placed near an object or a wall, a more detailed computation is done on that side. Firstly, we need to evenly subdivide the 3D model of the room such that each patch is small enough. Depending on the software used, the mesh of 3D room models typically comes with thousands of polygons in very small dimensions. However, in (Pelzer and Vorländer, 2010), it is found that any further surface subdivision below 70 cm by 70 cm in patch size is barely audible. Therefore we will have to group some of the polygons together to reduce the mesh resolution before auralizing it using our proposed method in 4. To explain our

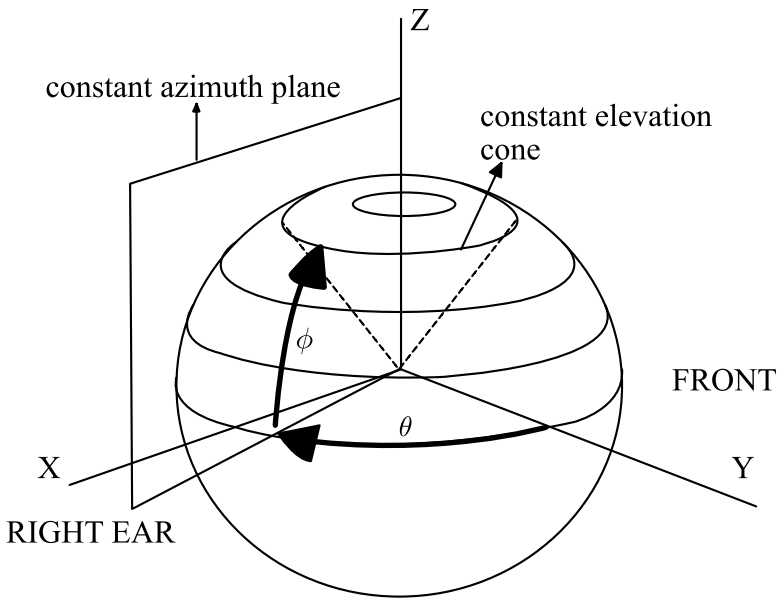

FIGURE 6.3: The vertical-polar coordinate system commonly used to describe head-related coordinate system in the literature. θ is azimuth angle, and ϕ is elevation angle.

method clearly, we label each surface polygon as p_i, $i = 1, ...K$ where K is the total number of polygons in the 3D mesh.

This brings us to the second step, where we have to create N_1 rays from from the listener's location in equal solid angle. To do this, we utilize Bauer's method (Bauer, 2000). This method is straightforward and it samples approximately evenly spaced point distribution on a sphere. Bauer proposed this method for stellar attitude determination analyses.

The algorithm presented in (Bauer, 2000) is as follows,

$$L = \sqrt{N_1 \pi} \tag{6.1}$$

$$z_k = 1 - \frac{2k-1}{N}, \qquad 1 \leq k \leq N_1 \tag{6.2}$$

$$\psi_k = \cos^{-1}(z_k), \qquad 1 \leq k \leq N_1 \tag{6.3}$$

$$\theta_k = L\psi_k, \qquad 1 \leq k \leq N_1 \tag{6.4}$$

$$x_k = \sin(\psi_k)\cos(\theta_k), \qquad 1 \leq k \leq N_1 \tag{6.5}$$

$$y_k = \sin(\psi_k)\sin(\theta_k), \qquad 1 \leq k \leq N_1 \tag{6.6}$$

$$\boldsymbol{d_k} = \{x_k, y_k, z_k\}^T, \tag{6.7}$$

where N_1 is the number of evenly distributed points on a unit circle with origin $\{0, 0, 0\}$. We can then shift the location of these points $\{x_k, y_k, z_k\}$ such that we end up

with a ray with its origin at listener's location L. We represent each ray as $\boldsymbol{r_k} = L + u \cdot \hat{\boldsymbol{d}}_{\boldsymbol{k}}$, where u is a scalar and $\hat{\boldsymbol{d}}_{\boldsymbol{k}}$ is a normalized vector $\boldsymbol{d_k}$. Each of the N rays $\boldsymbol{r}$ have its origin at the listener's location, and unit direction vector $\hat{\boldsymbol{d}}_{\boldsymbol{k}}$. Figure 6.4 illustrates 200 points generated at unit circle using Bauer's method (Bauer, 2000).

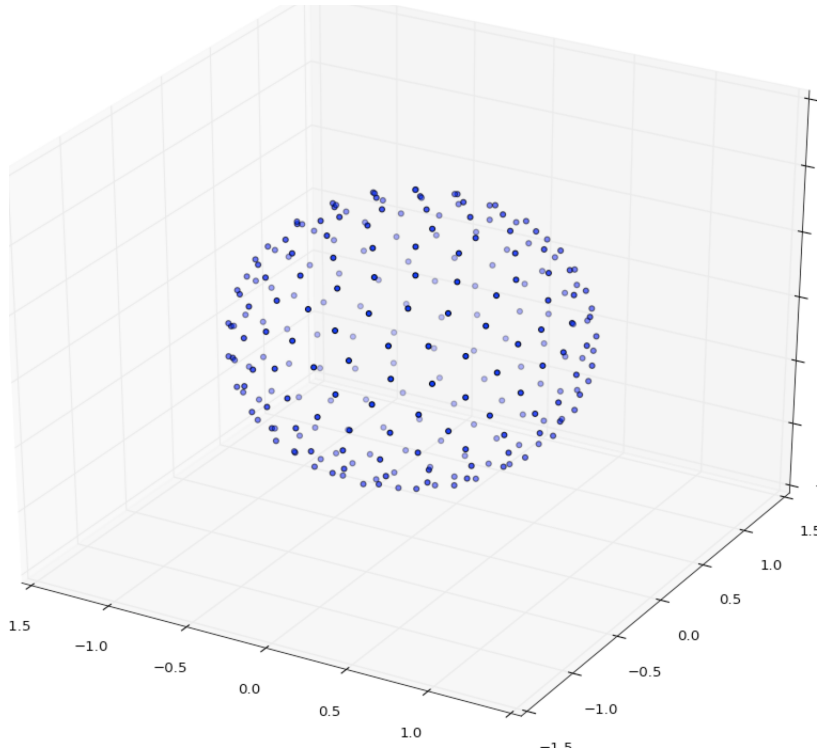

FIGURE 6.4: Illustration of generating 200 points at unit circle with approximately even solid angle using Bauer's method (Bauer, 2000).

The reflections from lateral directions are seen as more perceptually relevant (Kuttruff, 2009). The second step above guarantees that we can sample points almost evenly in a unit circle, but does not guarantee that we will sample points in the lateral direction. Therefore, the third step is to create another N_2 rays evenly in the listener's lateral direction (around the head, at the ear's height), parallel to the transverse plane. To generate N_2 points evenly in a unit circle around the listener and create its respective unit vector $\boldsymbol{d_l}$ one may simply do,

$$\phi_2 = \frac{360}{N_2}, \tag{6.8}$$

$$x_l = \cos(\phi_2 * l), \quad 1 \leq l \leq N_2, \tag{6.9}$$

$$y_l = \sin(\phi_2 * l), \quad 1 \leq l \leq N_2, \tag{6.10}$$

$$z_l = 0, \tag{6.11}$$

$$\boldsymbol{d_l} = \{x_l, y_l, z_l\}^T. \tag{6.12}$$

Similarly, we have N_2 lateral rays $\boldsymbol{r_l} = L + v \cdot \hat{\boldsymbol{d}}_{\boldsymbol{l}}$, where v is a scalar and $\hat{\boldsymbol{d}}_{\boldsymbol{l}}$ is normalized vector $\boldsymbol{d_l}$. In total, we have $N = N_1 + N_2$ rays $\boldsymbol{r_k}$, $k = 1, ..., N$. We name

these N_2 rays $\boldsymbol{r_l}$ as *lateral rays*. Figure 6.5 illustrates the generation of 200 points at unit circle using Bauer's method (Bauer, 2000) and 50 lateral points. The number of *channels* for HRTF (see Chapter 4 for details) is also set as N_2.

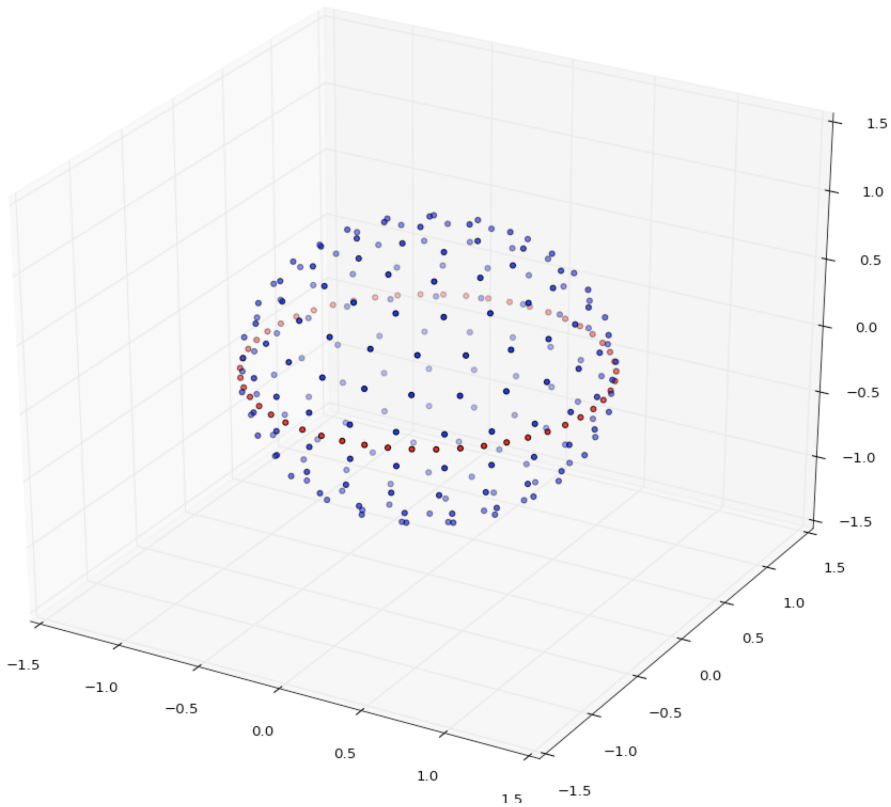

FIGURE 6.5: Illustration of generating 100 points at unit circle with approximately even solid angle using Bauer's method (Bauer, 2000), and 50 lateral points at unit circle.

Fourthly, we perform *ray casting* (Roth, 1982) to find the specific surface polygon p_i^* where each ray $\boldsymbol{r_k}$ intersects. We call p_i^* a *centroid* polygon. If we assume that the mesh is fine enough such that no two rays intersect the same polygon, then we end up with N centroid polygons p_i^*. The complexity of a naive algorithm to perform this operation is $O(MN)$. However we can make use of space-partitioning data structure such as the k-d tree (Bentley, 1975) to speed up the operation into $O(N \log M)$.

Fifthly, we group the surface polygons such that each group has exactly one centroid polygon p_i^*. We can do this by grouping the polygons based on the centroid polygon p_i^* with the nearest Eucledian distance. A naive algorithm has a complexity of $O(N^2)$, however it can be easily reduced to $O(N \log N)$ if k-d tree is used (Bentley, 1975). The reason for doing this is that ultimately, we will end up with N groups of surface polygons and we can use our method in Chapter 4 to model each group of surface patches using N FDN delay lines.

The final step is to apply our method in Chapter 4. The difference is that now we model the energy for each surface group in one delay line, instead of modeling each surface patch in one delay line. In other words, we sample monte carlo points from

all polygons that make up that each surface group. The length of the delay line is the amount of time taken for sound waves to travel from the source to the intersection point within that centroid polygon, and finally to the listener. The size of each surface groups may differ, and each delay line no longer represents acoustic energy from the same surface size.

It is possible to end up with a total distinct intersection polygons p_i^* that is less than N if two or more rays intersect the same polygon. However note that since all rays are distinct, they intersect at different points on the same polygon. In this case, we still proceed to the fourth step and compute the amount of energy from that surface group. Assume that a particular surface group has m centroid polygons. We then divide the amount of acoustic energy in this surface group by m, and model them individually using m delay lines in the FDN.

6.4 EVALUATION

6.4.1 BRIR Recordings

In this section we are going to evaluate our method with real BRIR recordings. We used BRIRs from the following rooms for our objective evaluation method:

- R1: A lift lobby (1.95m by 5.52m by 2.9m) in a basement. The material of the floors and walls is marble, and painted concrete for the ceiling. 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 entry-way 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 used four configurations of source and microphone positions (labeled P1, P2, P3 and P4) in R1 and R3 and three configurations (labeled P1, P2 and P3) in R2 for evaluation in this section. For BRIRs from (Jeub, Schäfe, and Var, 2009), we took two

configurations in R6 and one source-microphone configuration in each of the other rooms. In total, we used 15 BRIR recordings for the objective part of the evaluation.

Similar to the way we record BRIRs for evaluation in Chapter 4, to measure and record BRIRs in R1, R2, and R3, we used the logarithmic sine sweep method presented in (Farina, 2000). 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 (Hak, Wenmaekers, and Luxemburg, 2012). The response of the speaker is shown in Figure 4.2. 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 (Lundeby et al., 1995) 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.

6.4.2 BRIR Simulation

We simulated all 15 BRIRs using our method that we previously explained in Chapter 4, except that we now group the room patches based on the method we proposed in Section 6.3. The initial resolution for the room patches is 3750 surface meshes per room. The method is coded in C++ as an iOS app, and was run on an iPhone 8. The number of Monte-Carlo simulation per patch group is set to 100, and we simulated the BRIRs using FDN sizes of 16, 24, 32, 48, 64, 80, 96, 112 and 128. The number of rays N_2 in the lateral direction is set to 12, which is also equal to the number of channels set in the multiplexer. We also simulated the same 15 BRIRs using our method explained in Chapter4 as comparison, with the same set of FDN sizes 16, 24, 32, 48, 64, 80, 96, 112 and 128. The sizes of each surface patch is approximately constant. When the FDN size is not a power of two, we used the FDN design proposed in (Anderson et al., 2015), otherwise we used the original FDN proposed in (Jot and Chaigne, 1991) and Fast Hadamard Transform in place of the unitary mixing matrix. The number of points used in the Monte Carlo simulation is 100 per polygon for both methods.

6.4.3 Whiteness

In this section we briefly present the *whiteness* value of the proposed method. Using a lossless FDN of size 64 which includes 12 lateral rays, we obtained 15 sets of 131072 samples of BRIR. Note that this has the same settings as the BRIRs in section 6.4.2, except that these BRIRs are lossless (without decay). The BRIRs used for the next sections on objective and subjective evaluation are as per normal with the appropriate decay and reverberation time.

The 15 BRIRs settings have an average SFM value (2048 window size) of: 0.533, 0.555, 0.524, 0.499, 0.548, 0.545, 0.548, 0.536, 0.532, 0.537, 0.535, 0.568, 0.546, 0.567, and 0.568 respectively for all 15 samples. The averaged SFM of the entire 131072 samples for all 15 BRIRs is 0.543, which is higher (whiter) than the JND established in Chapter 2. Recall in Chapter 2 that we found the whiteness JND to be $\hat{Q} = 30.35$. As stated in chapter 3, the conversion from $\hat{Q}$ to SFM is,

$$\hat{\Xi} = e^{-(\hat{Q}*\frac{\sqrt{\pi^2/6-1}}{\sqrt{2048}}+\gamma)}, \tag{6.13}$$

where e is Euler's number and γ is Euler-Mascheroni constant. Therefore, using the equation above, we find that $\hat{\Xi} = 0.3288$. This shows that the proposed method does not colourise the FDN to the point that it is noticeable.

6.4.4 Objective Evaluation

Similar to Chapter 4, we evaluate our method based on six room acoustic parameters: IACC, D_{50}, C_{80}, T_S, RT_{60}, and EDT as suggested in ISO 3382-1:2009 (Iso3382-1, 2009). ISO 3381-1:2009 defines these room parameters to quantify the characteristics of a BRIR, measured in the 500Hz and 1000Hz frequency bands,

1. Reverberation time (RT_{60}), the time in seconds for the impulse response sound level to decay to 60dB below its initial value.

2. Early decay time (EDT), the time in seconds for the impulse response sound level to decay to 10dB below its initial level.

3. Definition (D_{50}), the ratio of the energy in the first 50ms of the BRIR to the total energy of the BRIR.

4. Clarity (C_{80}), the ratio of the energy in the first 80ms of the BRIR to the energy in the later part of the BRIR, measured in decibels.

5. Center Time (T_S), the center of gravity of the energy of the BRIR, measured in miliseconds.

6. Interaural Correlation Coefficient ($IACC_{E3}$), the measure of correlation of audio signals arriving between the left and right ears, also known as the 'spatial impression'. IACC is correlated with the apparent source width (ASW). $IACC_{E3}$ values come from interaural correlation coefficients between 0 and 80ms. Unlike the other parameters above, they are measured three octave bands: 500Hz, 1000Hz, and 2000Hz because it is suggested in (Hidaka, Beranek, and Okano, 1995) that these values can be used to directly represent the ASW.

All results except for IACC are averaged over the left and right channels of the BRIR.

To quantify the amount of error the simulated BRIRs has in terms of the above room parameters, we use the JND. As explained in the previous chapters, JND is defined as the smallest amount of change in a particular variable that is noticeable more than half of the subjects of interest (Fechner, 1966). Recall that the JND values for each of the six parameters described above are defined in (Iso3382-1, 2009) as follows,

1. RT_{60}: A deviation of 5% between measured and simulated values in the average of the 500Hz and 1000Hz frequency bands.

2. EDT: same as RT_{60}.

3. D_{50}: 0.05 absolute difference between measured and simulated values in the average of 500Hz and 1000Hz frequency bands.

4. C_{80}: 1 dB difference between measured and simulated values in the average of 500Hz and 1000Hz frequency bands.

5. T_S: 0.01 absolute second difference between measured and simulated values in the average of 500 and 1000Hz frequency bands.

6. $IACC_{E3}$: 0.075 absolute difference between measured and simulated values in the average of 500Hz, 1000Hz, and 2000Hz frequency bands.

We can directly set the reverberation time for our BRIR simulations such that it is below 0.5 JND of the recorded BRIR. Therefore we omit results on reverberation time.

Computation time

The time taken for our unoptimized code to perform any necessary parameter updates using the proposed method and $N = 16, 32, 64,$ and 128 are shown in table 6.1. Recall that we used the algorithm introduced in Chapter 4 to render the rest of the BRIRs, and that the method in Chapter 4 is an algorithmic reverb where it is able to directly process an input signal without the need to first produce a BRIR for later convolution. However we produced these BRIRs for the purpose of evaluation in this Chapter.

In our implementation, we used plain array data structure without any optimization. We used naive algorithm to find the centroid polygons and solve the nearest neighbour problem, i.e: group the rest of the polygons to the centroid polygon with the nearest Eucledian distance. The complexity for the naive nearest neighbour algorithm is $O(Nn^2)$, where N is the size of the FDN, and n is the number of the original polygons in the room, which is 3750.

The computational time to find the nearest centroid polygon can significantly be further sped up using spatial data structures such as the K-d tree (Bentley, 1975). The complexity of searching using K-d tree is guaranteed at $N \log_2 n$, where n is the number of points in the search space. This is a significant reduction from the complexity of the naive algorithm that we used. The K-d tree need to only be initialized once, with a complexity of $O(3n \log n)$, and is independent of listener or source location change at runtime as well as reusable each time the program that needs to render the BRIRs restart. By using the right data structure to optimize the nearest neighbour computation, it is very likely that the update time shown in Table 6.1 can be further reduced to be below $80ms$, which was suggested in () as the bench mark for real-time computation (Brungart, Simpson, and Kordik, 2005).

Results

Tables 6.2 to 6.4 present the average absolute JND of IACC, D_{50}, C_{80}, and T_S from all 15 BRIRs using 8, 12, and 16 lateral rays respectively. Similarly, Table 6.5 presents the average absolute JND of EDT from all 15 BRIRs using 8, 12, and 16 channels respectively. N represents the number of FDN delay lines, where smaller delay lines require lesser computational power at the cost of accuracy. The last column of Tables 6.2 to 6.5 contains the p-value result from Wilcoxon signed rank test, with the null hypothesis

TABLE 6.1: The time taken in miliseconds (ms) for our unoptimized code and naive algorithm to perform parameter updates using FDN sizes of $N = 16, 32, 64$ and 128.

	16	32	64	128
R1 P1	92.325	94.021	96.361	99.862
R1 P2	85.232	91.604	95.251	108.622
R1 P3	90.172	99.938	95.207	99.414
R1 P4	92.833	92.659	95.617	99.346
R2 P1	87.972	93.088	99.801	98.300
R2 P2	94.516	97.041	99.273	100.356
R2 P3	91.184	90.700	98.651	99.119
R3 P1	95.492	96.664	95.026	98.840
R3 P2	96.089	91.077	93.610	100.717
R3 P3	95.191	94.632	93.689	103.246
R3 P4	94.651	91.225	96.427	97.952
R4	92.690	91.366	98.331	98.485
R5	91.945	89.254	96.408	104.425
R6 P1	88.004	91.670	91.865	107.520
R6 P2	94.292	92.970	92.459	106.886
μ	92.173	93.194	95.865	101.539
σ	3.168	2.850	2.402	3.643

that the proposed method JND is less than the existing method JND at 5% significance level. Recall that the proposed method divides the room geometry as explained in Section 6.3 while the existing method uses an approximately even subdivision on all room surfaces.

The p-values that are lesser than 0.05 are printed in bold. The last column of tables 6.2 to 6.4 indicate that one can say with more than 95% certainty that in general, the proposed method improves the existing method in terms of all four room parameters when the FDN size is less than 64. However, we do not see the same amount of improvement for EDT. As previously mentioned in Chapter 4, the 3D model of the room may not be the same exact replica as the room in real life, and EDT is known to be highly affected by small errors introduced by the room modeling mesh itself (Iso3382-1, 2009).

An interesting observation from table 6.2 to 6.5 is that the JND from the proposed method seems to be almost around the same value regardless of the value of N. For example, in Table 6.2, the JND of IACC when $N = 16$ is 1.438, and it is slightly reduced to 1.204 when $N = 128$ using the proposed method, but we see a more dramatic fluctuation at 5.213 when $N = 16$ and 1.190 at $N = 128$ using the existing method. One explanation for this is due to the existence of the lateral rays. The lateral rays ensure that reflections from the lateral directions are always represented, regardless of the value of N.

Another informal observation from table 6.2 to 6.4 is that the errors from the proposed method seem to be reduced when the number of lateral rays are increased from 8

to 12. However the same improvement does not seem to be obvious when the number of lateral rays are further increased to 16. Therefore for the subjective evaluation in the next section (Section 6.4.5), we at used BRIRs that are simulated using the proposed method and 12 lateral rays.

While Table 6.2 to Table 6.5 summarize the averaged JND for all 15 BRIRs, Table 6.6 shows the absolute JND for all 15 simulated BRIRs and five room parameters using the proposed and existing method, with $N = 64$, and $N_2 = 12$ (12 lateral rays). In other words, Table 6.6 can be seen as the expansion of Table 6.3 on each row where $N = 64$. The JND values that are lesser than 1 are printed in bold. The last two rows of Table 6.6 presents the mean JND and standard deviation (σ) JND for all 15 BRIRs in terms of the respective room acoustic parameters. One direct observation from Table 6.6 is that both methods are able to render the BRIRs with less than 1 JND when evaluated in terms of the five room parameters (especially the first four: IACC, D_{50}, C_{80}, and T_S) more than half the time. This shows that when one looks at each of the 15 BRIRs the proposed method is comparable to the existing method in terms of objective evaluation performance, using FDN of size 64 and 12 lateral rays. However in the subjective evaluation, we present the result where the proposed method is superior to the existing method in terms of localization cues.

6.4.5 Subjective Evaluation

In this section, we present the result of our listening test from 19 candidates. All of the listening test candidates (12 males, 7 females, aged between 19 to 33) reported normal hearing condition. On average, the test took 45 minutes to complete and it was conducted in a small, carpeted, enclosed, and quiet meeting room with its air conditioning turned off. We encouraged each listener to take small breaks in between and we do not limit their time in completing the task to reduce fatigue. A pair of AKG-702 headphones, an amplifier, and iPad Air were used to playback the audio files.

Test Procedure

Four BRIRs were used (R1 P4, R2 P1, R4 and R5) for this listening test. Similar to the subjective test in Chapter 4, these BRIRs were selected such that we have a variation in room size, listener-source location, and reverberation time. Four 8s long anechoic input signals: a male spoken speech, a guitar piece (Woirgard et al., 2012), and a female spoken speech. We filtered frequencies below 100 Hz out from the audio files since geometric acoustics methods are known to not accurately reproduce the wave phenomena (diffraction and interference) when its wavelength is comparable to the size of everyday objects (Siltanen, Lokki, and Savioja, 2010).

Each listener was presented with three sets of audio files at a time. The first file was convolved with the measured BRIR, and the other two files were convolved with the simulated BRIRs using the proposed and existing method. The size of FDN used to simulate the BRIRs are 32, 64, and 128. The number of lateral rays used in the proposed method for this section is set to 12, and we used 100 Monte Carlo points per polygon for both methods. There are two tasks to do for each listener. Firstly, each listeners was asked to compare the degree of naturalness (less - more) of both simulated files to the measured files on a 15-point bipolar scale (anchored at -7 and 7 for both ends). As

previously mentioned in Chapter 4, this is a scale that is often used for subjective tests of perceptual qualities. Studies in (Chaiken and Eagly, 1983; South, Oltmanns, and Turkheimer, 2005) show that it can produce reliable results and reduce grade inflation. The descriptions for the ratings are: 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. Secondly, the listeners were tasked to identify the azimuthal direction of the sound source. The azimuth is sectioned into 14 sections, ranging from 0 to 7. Figure 6.6 illustrates the azimuthal direction. In other words, 0 means that the source direction in the simulated BRIR is the same as the source direction in the measured BRIR, 7 means that the source direction in the simulated BRIR is completely at the opposite direction, and ratings 1 to 6 are interpolated accordingly. In total for each task, the subjects listen to 36 (3 audio files per BRIR and FDN setting, four BRIRs, three FDN settings) of three sets of audio files.

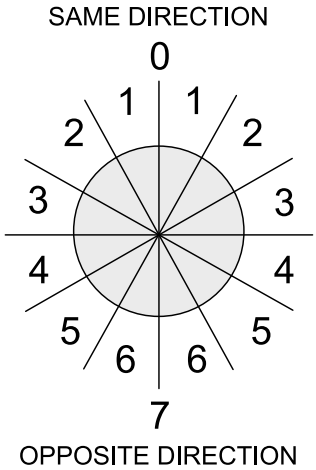

FIGURE 6.6: The azimuth direction for the ratings in the Localisation task of the listening test. 0 indicates the same direction of source in the measured BRIR as in the simulated BRIR, and 7 indicates the complete opposite source direction. Ratings 1 to 6 are interpolated accordingly and mirrored.

Results

Figure 6.7 shows the histograms of the listening test results. The left column shows the ratings from Naturalness, and the right column shows the ratings from Localisation. The ratings from the proposed method are represented by the darker bars, while the ratings from the existing method are represented by the lighter bars. The majority of the subjects feel that the simulated BRIRs using both methods are natural, as indicated by the bell-shaped histogram with a peak within ratings -1 to 1.

However the proposed method is superior than the existing method in terms of localization. From the histogram in Figure 6.7, we can see that more subjects give 0 and 1 ratings (exactly the same azimuth, or at least nearby) using the proposed method, as indicated by the taller dark bar at ratings 0 and 1, and taller gray bar when the ratings

grow higher (recall high ratings show higher discrepancy between the simulated and measured BRIRs).

Figure 6.8 and 6.9 show the boxplots of the listening test result (Localization and Naturalness) respectively. From the boxplots, we can see that median result for the proposed method is either comparable or lower than the existing method, for all FDN size settings of $N = 32, 64,$ and 128. From Figure 6.9, we can also deduce that both existing and proposed methods are reasonably natural for most test subjects, and that the size of N does not seem to affect much on the ratings. Similarly for localization, the size of N also does not seem to affect the ratings in any significant manner.

In summary, the most important takeaway from this listening test is that the proposed method is able to reproduce the localization cues more accurately as compared to the existing method. This is because the proposed method places more emphasis on surface points that are nearer to the listener ears, and also ensure that there is at least one computation point per channel. The proposed method addresses the problem shown in Figure 6.2, where not all HRTF azimuthal channels have an output, and therefore the localization cues from that current azimuthal section is lost.

6.5 SUMMARY AND FUTURE WORK

In this chapter, we introduced a way to group surface polygons in a virtual 3D room model such that its perceptual cues is still preserved. The method introduced in this chapter can be applied for any geometrical acoustics methods. To evaluate our method, we implemented the proposed way to group the surface polygons using our room acoustic rendering algorithm presented in Chapter 4, and we compared it to the results when we evenly group the room polygons equally (which we called *existing method*).

We presented both objective and subjective evaluation results. In the objective evaluation section, we conducted Wilcoxon signed-rank test to find out on whether the proposed method brings significant improvement to the existing method. The statistical test result shows that the proposed method brought significant improvement (0.05 significance level), depending on the values of N. In the case that there is no significant improvement, it does not bring about worse outcome. In the subjective evaluation section, we found that the proposed method seems to be as natural as the existing method, but it comes with better localization cues. The only drawback of the proposed method is the additional computational time needed to solve the nearest neighbour problem. However this can be easily fixed using efficient data structures such as the K-d tree (Bentley, 1975).

A potential area for future work is to further investigate what is the optimum number of lateral rays such that the listeners can no longer hear a difference. In this paper we only tested three cases of lateral rays: 8, 12, and 16, and we set the azimuthal channels for the HRTFs accordingly. We did not establish any pattern using these 3 settings, however further study to find out its saturation point may be useful in the case that computational power is limited.

TABLE 6.2: The average absolute JND values of IACC, D_{50}, C_{80}, and T_S for all 15 BRIRs simulated using proposed and existing method using various FDN size N and 8 lateral rays. The last column contains the p-value of Wilcoxon signed-rank test with the null hypothesis that the proposed method has less absolute JND than the existing method. p values that are lesser than 0.05 are printed in bold.

N	Prop. μ JND	Prop. σ JND	Exist. μ JND	Exist. σ	p value
			IACC		
16	1.438	1.284	5.213	2.110	**0.001**
24	1.241	1.077	2.820	2.029	**0.013**
32	1.329	0.734	2.163	1.801	0.063
48	1.456	0.997	1.545	1.290	0.476
64	1.154	1.044	1.896	1.484	0.087
80	1.315	1.040	1.813	1.364	0.198
96	1.125	1.087	1.589	1.244	0.127
112	1.158	0.807	1.335	1.089	0.326
128	1.204	0.911	1.190	1.185	0.456
			D_{50}		
16	0.815	0.709	5.293	3.463	**0.001**
24	0.836	0.756	2.838	2.432	**0.002**
32	1.134	0.953	2.782	2.366	**0.034**
48	0.704	0.723	1.798	1.969	**0.002**
64	0.681	0.899	1.383	1.413	**0.003**
80	0.893	0.795	1.354	1.393	0.087
96	0.654	0.702	1.297	1.778	0.127
112	0.884	0.729	1.379	1.769	0.212
128	0.804	0.674	1.257	1.607	0.248
			C_{80}		
16	0.592	0.468	9.949	6.402	**0.000**
24	0.865	0.676	2.955	3.217	**0.007**
32	1.026	0.875	2.927	2.702	**0.010**
48	0.785	0.648	2.265	1.519	**0.001**
64	0.818	0.653	0.961	0.802	0.140
80	0.980	0.686	1.033	0.832	0.390
96	0.727	0.616	1.223	1.253	**0.027**
112	0.768	0.550	1.034	1.263	0.284
128	0.700	0.567	0.956	1.087	0.166
			T_S		
16	0.798	0.840	5.107	3.424	**0.000**
24	0.934	0.876	2.726	2.292	**0.007**
32	0.987	1.171	2.494	2.164	**0.010**
48	0.921	0.878	1.864	1.894	**0.007**
64	1.005	1.087	1.324	1.363	0.087
80	1.012	0.922	1.282	1.396	0.087
96	0.864	0.843	1.268	1.682	0.166
112	0.990	0.910	1.225	1.590	0.390
128	0.910	0.879	1.179	1.489	0.345

TABLE 6.3: The average absolute JND values of IACC, D_{50}, C_{80}, and T_S for all 15 BRIRs simulated using proposed and existing method using various FDN size N and 12 lateral rays. The last column contains the p-value of Wilcoxon signed-rank test with the null hypothesis that the proposed method has less absolute JND than the existing method. p values that are lesser than 0.05 are printed in bold.

N	Prop. μ JND	Prop. σ JND	Exist. μ JND	Exist. σ	p value
			IACC		
16	1.045	0.918	2.806	1.776	**0.005**
24	1.030	0.861	3.436	2.017	**0.021**
32	1.040	1.069	2.591	1.795	**0.015**
48	0.994	1.267	1.840	1.386	0.063
64	1.039	1.151	1.619	0.865	**0.049**
80	0.855	0.897	1.461	1.093	0.063
96	1.010	1.098	1.535	1.440	0.117
112	1.060	1.134	1.384	1.307	0.117
128	0.806	0.715	1.392	1.191	0.078
			$\mathbf{D_{50}}$		
16	0.757	0.822	1.604	1.672	**0.027**
24	0.831	0.807	2.878	2.425	**0.045**
32	0.883	0.838	2.516	2.411	**0.008**
48	0.883	0.745	1.773	2.010	**0.015**
64	0.794	0.860	1.457	1.359	**0.001**
80	0.782	0.661	1.585	1.259	**0.005**
96	0.805	0.782	1.277	1.785	0.284
112	0.679	0.697	1.376	1.791	0.166
128	0.731	0.738	1.248	1.565	0.181
			$\mathbf{C_{80}}$		
16	1.498	0.859	2.132	1.706	0.212
24	0.709	0.704	3.039	3.239	**0.021**
32	0.957	0.782	3.045	2.698	**0.006**
48	0.888	0.768	2.280	1.567	**0.002**
64	0.938	0.857	0.948	0.815	0.409
80	0.627	0.590	1.022	0.925	0.166
96	0.826	0.649	1.238	1.240	0.087
112	0.795	0.645	0.946	1.241	0.390
128	0.662	0.549	0.920	0.965	0.069
			$\mathbf{T_S}$		
16	0.918	0.883	1.658	1.094	**0.005**
24	0.782	0.891	2.752	2.291	**0.027**
32	0.832	1.028	2.476	2.185	**0.002**
48	1.035	1.179	1.895	1.900	**0.012**
64	0.907	1.080	1.279	1.343	**0.013**
80	1.126	0.779	1.154	1.323	0.433
96	0.992	0.823	1.229	1.677	0.390
112	0.986	0.925	1.226	1.566	0.284
128	1.027	0.941	1.173	1.431	0.433

TABLE 6.4: The average absolute JND values of IACC, D_{50}, C_{80}, and T_S for all 15 BRIRs simulated using proposed and existing method using various FDN size N and 16 lateral rays. The last column contains the p-value of Wilcoxon signed-rank test with the null hypothesis that the proposed method has less absolute JND than the existing method. p values that are lesser than 0.05 are printed in bold.

N	Prop. μ JND	Prop. σ JND	Exist. μ JND	Exist. σ	p value
			IACC		
16	1.372	0.964	5.557	2.111	**0.001**
24	1.008	0.982	3.542	2.029	**0.002**
32	0.928	1.092	2.658	1.813	**0.005**
48	0.778	1.087	1.967	1.309	**0.003**
64	1.132	1.021	1.912	1.295	0.056
80	1.112	1.054	1.757	1.341	**0.049**
96	0.898	1.057	1.585	1.332	0.063
112	1.193	0.992	1.443	1.275	0.364
128	0.949	0.925	1.569	1.249	0.056
			$\mathbf{D_{50}}$		
16	0.658	0.618	5.326	3.485	**0.001**
24	0.875	0.740	2.868	2.425	**0.001**
32	0.936	0.805	2.747	2.375	**0.007**
48	0.718	0.798	1.761	1.957	**0.002**
64	0.747	0.850	1.387	1.424	**0.008**
80	0.652	0.682	1.398	1.418	**0.002**
96	0.538	0.805	1.275	1.756	**0.018**
112	0.670	0.675	1.347	1.774	0.087
128	0.758	0.668	1.254	1.604	0.078
			$\mathbf{C_{80}}$		
16	0.711	0.556	9.938	6.402	**0.000**
24	0.744	0.559	2.985	3.251	**0.007**
32	0.735	0.672	2.933	2.713	**0.003**
48	0.651	0.668	2.212	1.511	**0.001**
64	0.837	0.657	0.960	0.829	0.268
80	0.694	0.678	1.125	0.822	**0.031**
96	0.693	0.713	1.261	1.257	**0.010**
112	0.781	0.648	0.975	1.228	0.364
128	0.683	0.619	1.013	1.064	0.117
			$\mathbf{T_S}$		
16	0.770	0.769	5.115	3.463	**0.000**
24	0.873	0.937	2.738	2.295	**0.003**
32	1.023	0.943	2.482	2.165	**0.006**
48	1.016	1.073	1.869	1.876	**0.012**
64	0.882	1.036	1.358	1.377	**0.023**
80	1.037	0.919	1.327	1.396	0.117
96	0.882	1.028	1.216	1.673	0.127
112	0.941	0.876	1.228	1.573	0.390
128	1.062	0.827	1.141	1.483	0.268

TABLE 6.5: The average absolute JND values of EDT for all 15 BRIRs simulated using proposed and existing method using various FDN size N and 8, 12, and 16 lateral rays. The last column contains the p-value of Wilcoxon signed-rank test with the null hypothesis that the proposed method has less absolute JND than the existing method. p values that are lesser than 0.05 are printed in bold.

N	Prop. μ JND	Prop. σ JND	Exist. μ JND	Exist. σ	p value
			$N_2 = 8$		
16	2.852	2.690	8.247	8.367	**0.002**
24	3.368	3.467	5.102	6.621	0.433
32	2.766	2.153	4.809	5.700	0.230
48	2.530	2.480	3.285	2.898	0.056
64	3.149	2.257	3.122	4.728	0.063
80	3.538	3.978	3.197	4.474	0.230
96	3.018	2.521	2.670	2.263	0.117
112	2.938	3.115	2.710	2.253	0.326
128	2.587	1.981	2.342	2.077	0.248
			$N_2 = 12$		
16	2.824	2.341	3.652	3.061	0.127
24	2.754	2.192	5.064	6.641	0.390
32	2.802	2.200	4.928	5.895	0.230
48	3.037	2.691	3.358	2.913	0.390
64	3.476	2.329	3.731	5.720	0.154
80	2.614	2.181	3.392	5.260	0.456
96	2.785	2.463	2.669	2.285	0.248
112	3.009	2.654	2.718	2.167	0.212
128	2.917	2.102	2.352	2.245	0.118
			$N_2 = 16$		
16	2.569	2.787	8.176	8.417	**0.001**
24	3.296	3.426	5.241	6.802	0.345
32	2.759	2.552	4.831	5.851	0.069
48	2.893	3.208	3.242	2.874	0.198
64	3.087	1.971	3.067	4.593	0.212
80	2.748	2.417	3.098	4.303	0.364
96	2.859	1.885	2.816	2.371	0.500
112	3.048	2.055	2.704	2.143	0.230
128	2.648	2.225	2.403	2.237	0.212

TABLE 6.6: FDN 64, 12 lateral rays

Room	IACC Prop.	IACC Exst.	D50 Prop.	D50 Exst.	C80 Prop.	C80 Exst	TS Prop.	TS Exst	EDT Prop.	EDT Exst.
R1 P1	**0.14**	1.26	1.31	1.46	**0.57**	**0.66**	**0.32**	**0.19**	1.43	2.19
R1 P2	**0.93**	1.30	**0.92**	1.31	**0.49**	**0.50**	1.77	2.01	2.10	1.25
R1 P3	1.62	**0.82**	**0.41**	1.43	**0.08**	**0.19**	**0.29**	1.32	1.41	1.53
R1 P4	**0.42**	3.77	2.46	5.33	2.79	2.94	3.36	5.03	4.18	2.35
R2 P1	**0.30**	1.71	**0.46**	2.46	1.45	1.10	**0.75**	**0.40**	4.09	4.65
R2 P2	4.48	2.60	2.01	2.30	1.98	1.47	**0.14**	**0.17**	6.30	6.53
R2 P3	**0.57**	**0.96**	**0.22**	**0.64**	1.06	**0.79**	2.00	1.64	5.01	5.87
R3 P1	2.52	1.14	2.49	2.98	**0.81**	0.48	2.93	3.42	2.08	1.04
R3 P2	**0.39**	1.05	**0.41**	**0.98**	**0.09**	**0.19**	**0.19**	1.14	1.40	**0.31**
R3 P3	1.10	3.09	**0.03**	**0.64**	**0.16**	**0.62**	**0.11**	1.00	1.22	**0.90**
R3 P4	**0.45**	1.67	**0.24**	**0.82**	1.95	2.03	**0.69**	**0.84**	**0.06**	**0.00**
R4	1.37	1.07	**0.35**	**0.19**	**0.08**	**0.03**	**0.32**	**0.67**	5.04	2.66
R5	**0.76**	1.35	**0.08**	**0.28**	**0.74**	1.94	**0.50**	**0.37**	8.47	23.08
R6 P1	**0.43**	1.58	**0.29**	**0.70**	1.70	**0.46**	**0.04**	**0.43**	3.95	3.47
R6 P2	**0.11**	**0.91**	**0.23**	**0.32**	**0.11**	**0.83**	**0.21**	**0.55**	5.42	**0.14**
μ	1.04	1.62	**0.79**	1.46	**0.94**	**0.95**	**0.91**	1.28	3.48	3.73
σ	1.15	**0.86**	**0.86**	1.36	**0.86**	**0.81**	1.08	1.34	2.33	5.72

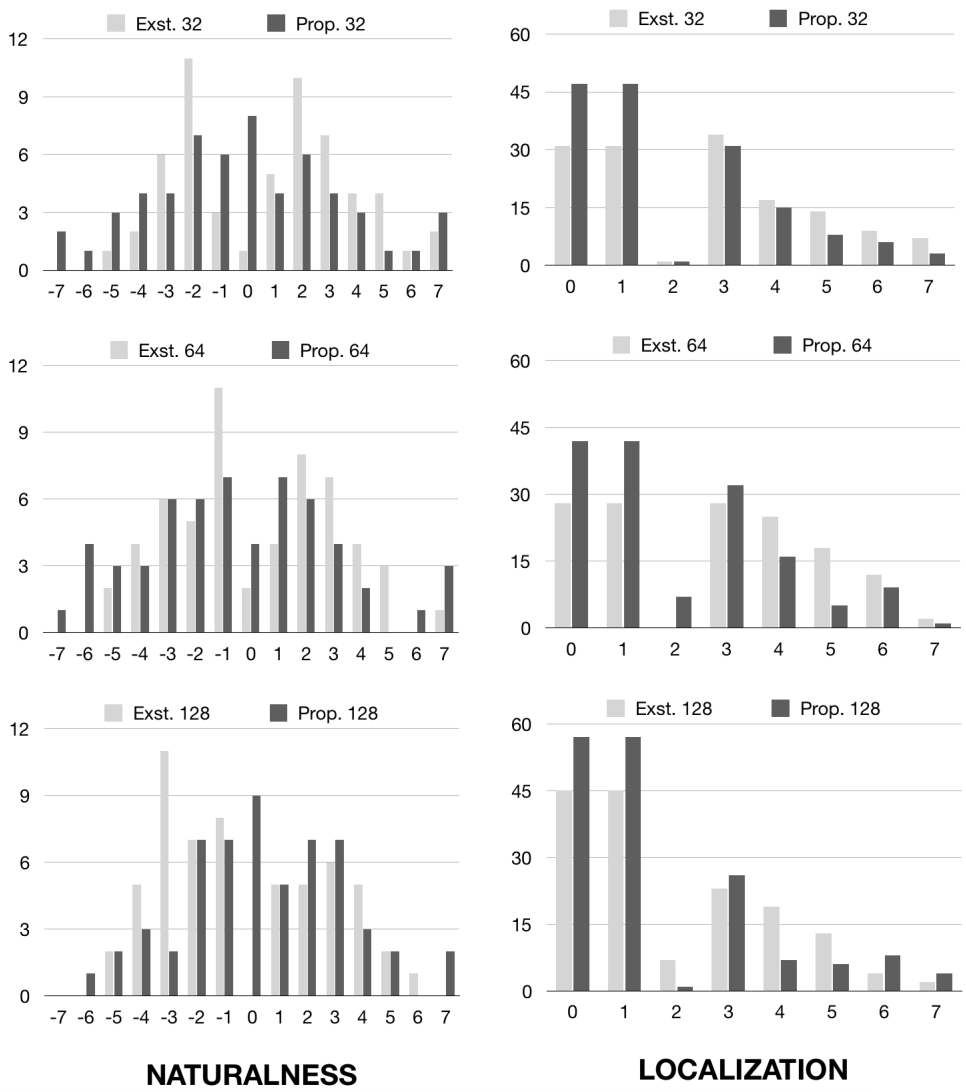

FIGURE 6.7: Listening test results on the ratings of Naturalness and Localisation from all 19 subjects.

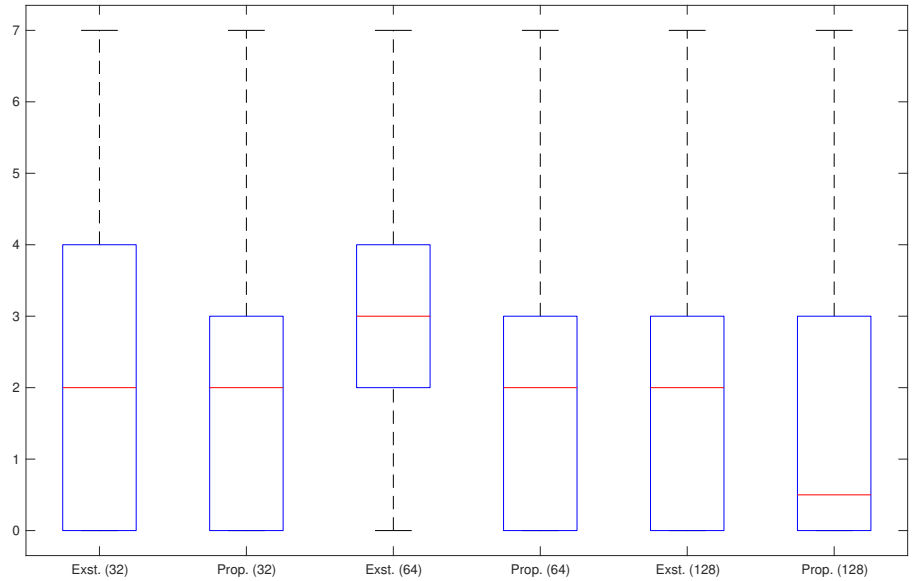

FIGURE 6.8: Boxplots of test results on the ratings of Localisation from all 19 subjects.

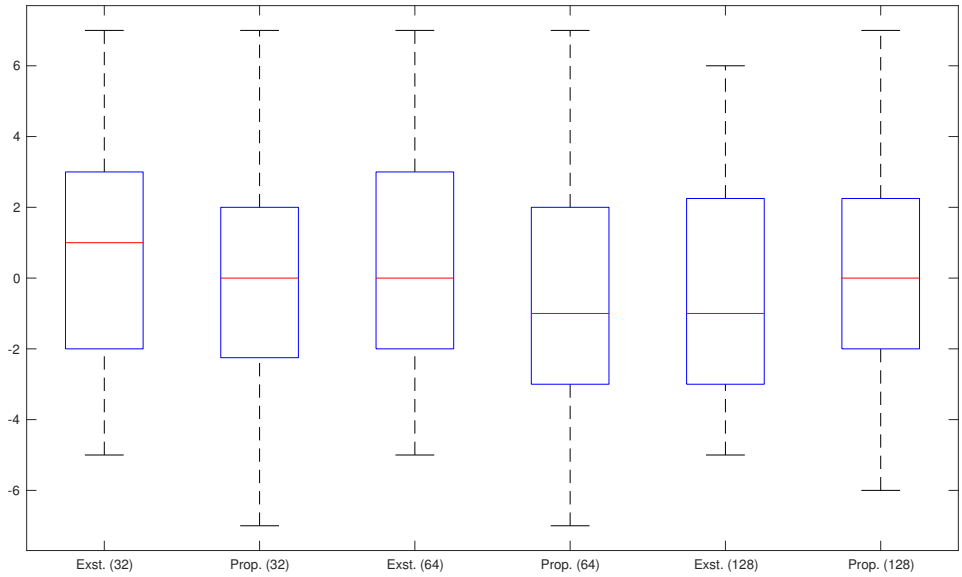

FIGURE 6.9: Boxplots of test results on the ratings of Naturalness from all 19 subjects.

Chapter 7

Conclusion

In section 1.2 we stated that we aim to develop a minimally efficient binaural room simulation algorithm design that it is able to directly auralize input signals in real time without the need of convolution, and yet remains perceptually convincing. Suitable applications include virtual reality plug-ins, music and film processing, and video games on both mobile and desktop. These are the applications where perceptual plausibility and computational load are prioritized over numerical accuracy.

Therefore in Chapter 4, we presented a method that is able to auralize dry input signals in real-time, even on mobile devices, with reasonable degree of perceptual plausibility. Both objective and subjective evaluation results were shown to prove that the method proposed is able to realistically auralize input signals. Also in the following Chapter 5, we expanded our method such that the same principles can be applied to existing hybrid auralization algorithms, and that these existing models can benefit from the principles established in Chapter 1 and 4.

The foundation for the method in Chapter 4 and 5 is presented in the beginning of this work in Chapter 1. This includes the establishment of physical significance of audio signals in FDN, the ray-tracing delay lines, the ARE, and the basic theories behind BRIRs and FDN itself. However since the FDN is highly affected by the composition of its delay lines length, we need to first investigate whether the ray-tracing delay lines will introduce unwanted artifacts on the output of lossless FDN, which ideally is supposed to be white or flat. In particular, what we need to find is the maximum amount of spectral variance (threshold, or JND) that can be present before a supposedly white output signal from the lossless FDN is perceived as colored. Since there is no such threshold (or JND) in the literature, we proceeded by studying the JND for spectral variance in Chapter 2.

Afterwards, in Chapter 3, we used the JND value found in Chapter 2 and showed that although there is some coloration in the FDN lossless output when we use the ray-tracing delay lines, this amount of coloration is still below the spectral variance JND and hence it is very likely that this coloration is not noticeable. This means that it is reasonable to implement ray-tracing delay lines in our method.

As the final part of our work, in Chapter 6, we offered a new method to group the polygons of a virtual room 3D model in a way that improves localization cues in the lateral plane, as well as improve the rendered BRIR accuracy in terms of room acoustic parameters (objective evaluation). Initially, when implementing and evaluating the methods offered in Chapter 4 and 5, we grouped polygons in the room into roughly equal amounts or area. We showed how this brought about problems when a listener is placed near a surface geometry, and that some cues from the lateral directions will

be lost. The method in Chapter 6 is set to address and minimize this problem. This polygon grouping method in Chapter 6 is applicable for the methods in Chapter 4 and 5, or any other geometrical acoustic methods that need to perform acoustic energy computations on virtual room surfaces.

We also presented possible directions for future work at the end of each chapter. In summary, for the work in Chapter 2, it may be worthwhile to investigate the JND for spectral flatness based on octave frequency bands, which is known to be more related to the perception of human auditory system. For the work in Chapter 3, it may be useful to collate all know methods of setting FDN delay lengths and investigate if the whiteness of its lossless output is below the JND established in Chapter 2 as a way to evaluate if a particular setting is better (in the sense of spectral flatness) than the other. In Chapter 4 and 5, it would be worthwhile to look into further modifications such that these methods are also applicable for other artificial reverberators, such as the Scattering Delay Network (Sena et al., 2015) or the circulant structure introduced in (Anderson et al., 2015). For the work in Chapter 6, it is possible to look into the effects of various amount of lateral rays, and establish an optimum amount to minimize computational power without compromising its perceptual quality.

Bibliography

Agus, N. et al. (2017). *Energy-Based Binaural Acoustic Modeling*. Tech. rep. 1. https://istd.sutd.edu.sg/research /technical-reports/energy-based-binaural-acoustic-modeling. Singapore University of Technology and Design.

Allen, J. B. and D. A. Berkley (1979). "Image method for efficiently simulating small room acoustics". In: *The Journal of the Acoustical Society of America* 64.943. https://doi.org/10.1121/1.382599, pp. 943–950. URL: https://www.google.com.sg/search?client=safari&rls=en&q=image+source&ie=UTF-8&oe=UTF-8&gws_rd=cr&ei=CryxWJihC4rKvgSt8Y7oBw#q=image+method+allen&*.

Anderson, H. et al. (2015). "Flatter Frequency Response from Feedback Delay Network Reverbs". In: *41st International Computer Music Conference 2015*. Denton, Texas. URL: http://quod.lib.umich.edu/i/icmc/bbp2372.2015.048/--flatter-frequency-response-from-feedback-delay-network?view=text&seq=4&size=150.

Anderson, H; et al. (2017). "Modeling the Proportion of Early and Late Energy in Two-Stage Reverberators". In: *Journal of the Audio Engineering Society* 65.12. https://doi.org/10.17743/jaes.2017.0041, pp. 1071–1031.

Angel, E. J., V. R. Algazi, and R. O. Duda (2002). "On the design of canonical sound localization environments". In: *113th Audio Engineering Society Convention*. 5714. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.139.6021.

Antani, Lakulish and Dinesh Manocha (2013). "Aural proxies and directionally-varying reverberation for interactive sound propagation in virtual environments." In: *IEEE transactions on visualization and computer graphics*. Vol. 19. 4. http://dx.doi.org/ 10.1109/TVCG.2013.27, pp. 567–575. URL: http://view.ncbi.nlm.nih.gov/pubmed/23428440.

Arntzen, M., L. Bertsch, and D. G. Simons (2015). "Auralization of novel aircraft configurations". In: *5th CEAS Air and Space conference*. Delft (NL). URL: http://repository.tudelft.nl/islandora/object/uuid:79d9c4d6-6422-47fb-933f-faf43f16e44a?collection=research.

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

Ballachanda, B. B. (1997). "Theoretical and applied external ear acoustics." In: *Journal of the American Academy of Audiology* 8.6, pp. 411–420. ISSN: 1050-0545. URL: http://view.ncbi.nlm.nih.gov/pubmed/9433687.

Barron, M. and A. H. Marshall (1981). "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure". In: *Journal of Sound and Vibration* 77.2. https://doi.org/10.1016/S0022-460X(81)80020-X, pp. 211–232. URL: https://doi.org/10.1016/S0022-460X(81)80020-X.

Bauer, Robert (2000). "Distribution of Points on a Sphere with Application to Star Catalogs". In: *Journal of Guidance, Control, and Dynamics* 23.1, pp. 130–137. ISSN: 0731-5090. DOI: 10.2514/2.4497. URL: http://dx.doi.org/10.2514/2.4497.

Bentley, Jon L. (1975). "Multidimensional binary search trees used for associative searching". In: *Communications of the ACM* 18.9, pp. 509–517. ISSN: 00010782. DOI: 10.1145/361002.361007. URL: http://dx.doi.org/10.1145/361002.361007.

Brown, C. P. and R. O. Duda (1998). "A structural model for binaural sound synthesis". In: *IEEE Transactions on Speech and Audio Processing* 6.5. http://dx.doi.org/10.1109/89.709673, pp. 476–488. ISSN: 1063-6676. DOI: 10.1109/89.709673. URL: http://dx.doi.org/10.1109/89.709673.

Brungart, D. S., B. D. Simpson, and A. J. Kordik (2005). "The detectability of headtracker latency in virtual audio displays". In: *International Conference on Auditory Display*. Vol. 73.

Buck, Adam et al. (2012). "Measurements of the just noticeable difference for reverberation time using a transformed up–down adaptive method". In: *The Journal of the Acoustical Society of America* 132.2060. https://doi.org/10.1121/1.4755592.

Carlile, S. (2013). *Virtual Auditory Space: Generation and Applications*. Springer Science & Business Media, pp. 42+.

Carty, B. and V. Lazzarini (2010). "Hrtf-early and Hrtf-reverb: Flexible Binaural Reverberation Processing". In: *International Computer Music Association*. New York. URL: http://quod.lib.umich.edu/i/icmc/bbp2372.2010.105/--hrtfearly-and-hrtfreverb-flexible-binaural-reverberation?view=image.

Chaiken, S. and A. H. Eagly (1983). "Communication modality as a determinant of persuasion: The role of communicator salience." In: *Journal of Personality and Social Psychology* 45.2, pp. 241–256. ISSN: 0022-3514. DOI: 10.1037/0022-3514.45.2.241. URL: http://dx.doi.org/10.1037/0022-3514.45.2.241.

Cornsweet, T. N. (1962). "The Staircase-Method in Psychophysics". In: *The American Journal of Psychology* 75.3, pp. 485–491.

Cox, H. (1966). "Linear versus Logarithmic Averaging". In: *The Journal of the Acoustical Society of America* 39.4, pp. 688–690. URL: http://asa.scitation.org/doi/abs/10.1121/1.1909942.

Crocker, M. J. (2007). *Handbook of Noise and Vibration Control*. John Wiley & Sons, pp. 13+.

Dahl, L. and J. M. Jot (2000). "A reverberator based on absorbent all-pass filters". In: *DAFx-00: Proceedings*. Verona, Italy.

Davis, Gary D. and Ralph Jones (1989). *The Sound Reinforcement Handbook*. Hal Leonard Publishing Coorporation, pp. 30+. ISBN: 0881889008.

Dougherty, Geoff (2009). *Digital Image Processing for Medical Applications*. Cambridge University Press, pp. 251+. ISBN: 0521860857, 9780521860857.

Drouiche, K. (2000). "A New Test for Whiteness". In: *IEEE Trans. on Signal Proc.* 48.7, pp. 1864–1871.

Duda, R. O., V. R. Algazi, and D. M. Thompson (2002). "The Use of Head-and-Torso Models for Improved Spatial Sound Synthesis". In: *113 Audio Engineering Society Convention*. 5712. URL: http://www.aes.org/e-lib/browse.cfm?elib=11294.

Duifhuis, H. (1973). "Audibility of Harmonics in Periodic " White Noise"". In: *The Journal of the Acoustical Society of America* 54.1, p. 316. ISSN: 0001-4966. DOI: 10.1121/1.1978255. URL: http://dx.doi.org/10.1121/1.1978255.

Farina, A. (2000). "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique". In: *108th Convention of the Audio Engineering Society*. 5063. URL: http://www.aes.org/e-lib/browse.cfm?elib=10211.

Fechner, G. (1966). "Elements of psychophysics". In: ed. by H. Adler. Vol. 1. Translation of "Elemente der Psychophysik". New York: Holt Rinehart Winston.

Funkhouser, T., J. M. Jot, and N. Tsingos (2002). *Sounds good to me!* URL: https://www.siggraph.org/s2002/conference/courses/crs45.html.

Gerzon, Michael A. (1973). "Periphony: With-Height Sound Reproduction". In: *Journal of the Audio Engineering Society* 21.1, pp. 2–1–.

Gescheider, George A. (1997). *Psychophysics: The Fundamentals*. Third. Psychology Press. Chap. 3.

Gray, A. and J. Markel (1974). "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 22.3, pp. 207–217. URL: http://ieeexplore.ieee.org/document/1162572/.

Griesinger, D. (1996). "Spaciousness and envelopment in musical acoustics". In: *Proceedings of 101st Audio Engineering Society Convention*. 4401. Los Angeles, CA. URL: http://www.aes.org/e-lib/browse.cfm?elib=7378.

— (1999). "The Science of Surround". In: 3 Oak Park, Bedford, MA 01730.

— (2000). "Creating Reverb Algorithms For Surround Sound". In: https://web.archive.org/web/20140914084732 /http://www. soundonsound.com/sos/mar00/ articles/dave.htm. Sound on Sound. URL: https://www.google.com.sg/search?client=safari&rls=en&q=Creating+Reverb+Algorithms+For+Surround+Sound&ie=UTF-8&oe=UTF-8&gfe_rd=cr&ei=VzdiV6WiE9HbugSk_bOABQ.

— (2010). "The relationship between audience engagement and the ability to perceive pitch, timbre, azimuth and envelopment of multiple sources". In: *Proceedings of the International Symposium on Room Acoustics, ISRA 2010*. Melbourne, Australia.

Guilford, J. P. (1954). *Psychometric Methods*. McGraw- Hill, New York.

Guttman, N. and B. Julesz (1963). "Lower limits of auditory periodicity analysis". In: *Journal of the Acoustical Society of America* 35. https://doi.org/10.1121/1.1918551, pp. 610+.

Haas, Helmut (1972). "The Influence of a Single Echo on the Audibility of Speech". In: *Journal of the Audio Engineering Society* 20.2, pp. 146–159.

Hak, C. C. J. M., R. H. C. Wenmaekers, and L. C. J. van Luxemburg (2012). "Measuring Room Impulse Responses: Impact of the Decay Range on Derived Room Acoustic Parameters". In: *Acta Acustica united with Acustica* 98.6. https://doi.org/10.3813/AAA.918574, pp. 907–915. ISSN: 1610-1928. DOI: 10.3813/aaa.918574. URL: http://dx.doi.org/10.3813/aaa.918574.

Hampel, F. R. (1974). "The Influence Curve and Its Role in Robust Estimation". In: *Journal of the American Statistical Association* 69.346, pp. 383–393.

Harma, A. and U. K. Laine (2001). "A comparison of warped and conventional linear predictive coding". In: *IEEE Transactions on Speech and Audio Processing* 9.5, pp. 579–588.

Harris, L. E. and K. R. Holland (2009). "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*. Vol. 31. 4. Brighton, UK, pp. 294–309.

Hellstrom, P. A. and A. Axelsson (1993). "Miniature microphone probe tube measurements in the external auditory canal." In: *The Journal of the Acoustical Society of America* 93.2, pp. 907–919. ISSN: 0001-4966. URL: http://view.ncbi.nlm.nih.gov/pubmed/8445126.

Hidaka, T., L. L. Beranek, and T. Okano (1995). "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls". In: *The Journal of the Acoustical Society of America* 98.2. http://dx.doi.org/10.1121/ 1.412847, pp. 988–1007. ISSN: 0001-4966. DOI: 10.1121/1.414451. URL: http://dx.doi.org/10.1121/1.414451.

Hidaka, Takayuki, Yoshinari Yamada, and Takehiko Nakagawa (2007). "A new definition of boundary point between early reflections and late reverberation in room impulse responses." In: *The Journal of the Acoustical Society of America* 122.1. http://dx.doi.org/10.1121/ 1.2743161, pp. 326–332. ISSN: 1520-8524. DOI: 10.1121/1.2743161. URL: http://dx.doi.org/10.1121/1.2743161.

Iso3382-1 (2009). "ISO 3382-1:2009 - Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces". In: URL: http://www.iso.org/iso/catalogue_detail.htm?csnumber=40979.

Jeub, M., M. Schäfe, and P. Var (2009). "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. https://doi.org/ 10.1109/ICDSP.2009.5201259. Santorini, Greece: IEEE Press, pp. 550–554. ISBN: 978-1-4244-3297-4. URL: http://portal.acm.org/citation.cfm?id=1700398.

Johnston, J. D. (1988). "Transform coding of audio signals using perceptual noise criteria". In: *IEEE Journal on Selected Areas in Communications* 6.2, pp. 314–323. ISSN: 07338716. DOI: 10.1109/49.608. URL: http://dx.doi.org/10.1109/49.608.

Jot, J. M. (1997). "Efficient Models for Reverberation and Distance Rendering in Computer Music and Virtual Audio Reality". In: *Proc. 1997 International Computer Music Conference*. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.8176.

Jot, J. M. and A. Chaigne (1991). "Digital Delay Networks for Designing Artificial Reverberators". In: *90th Convention of the Audio Engineering Society*. 3030. URL: http://www.aes.org/e-lib/browse.cfm?elib=5663.

Kaernbach, C. (2001). "Parameters of echoic memory". In:

Kajiya, James T. (1986). "The rendering equation". In: *Proceedings of the 13th annual conference on Computer graphics and interactive techniques - SIGGRAPH '86*. Vol. 20.

SIGGRAPH '86 4. Not Known: ACM Press, pp. 143–150. ISBN: 0897911962. DOI: 10.1145/15922.15902. URL: http://dx.doi.org/10.1145/15922.15902.

Kiminki, S. (2005). "Sound Propagation Theory for Linear Ray Acoustic Modelling". http://www.niksula.hut.fi/~skiminki/D- skiminki.pdf. MA thesis. 02150 Espoo, Finland: Helsinki University of Technology. Chap. 3.2.5.

Kuttruff, Heinrich (2009). "Room Acoustics, Fifth Edition". In: 5th ed. CRC Press. Chap. 5.1, pp. 131–132. ISBN: 0415480213. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0415480213.

Lehmann, E. A. and A. M. Johansson (2010). "Diffuse Reverberation Model for Efficient Image-Source Simulation of Room Impulse Responses". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.6. https://doi.org/10.1109/TASL.2009.2035038, pp. 1429–1439. ISSN: 1558-7916. DOI: 10.1109/tasl.2009.2035038. URL: http://dx.doi.org/10.1109/tasl.2009.2035038.

Lentz, T. (2007). "Binaural technology for virtual reality". PhD thesis. Aachen, Germany: RWTH Aachen University. URL: http://publications.rwth-aachen.de/record/51299.

Lindau, A. (2015). *Spatial Audio Quality Inventory (SAQI). Test Manual. v1.2*. Tech. rep. TU Berlin.

Litovsky, R. Y. et al. (1999). "The precedence effect". In: *The Journal of the Acoustical Society of America* 106.4 Pt 1. http://dx.doi.org/10.1121/1.427914, pp. 1633–1654. ISSN: 0001-4966. URL: http://view.ncbi.nlm.nih.gov/pubmed/10530009.

Ljung, G. M. and G. E. P. Box (1978). "On a Measure of Lack of Fit in Time Series Models". In: *Biometrika* 65.2, pp. 297–303. URL: http://stat.wharton.upenn.edu/~steele/Courses/956/Resource/TestingNormality/LjungBox.pdf.

Lundeby, A. et al. (1995). "Uncertainties of Measurements in Room Acoustics". In: *Acta Acustica united with Acustica*, pp. 344–355. ISSN: 1610-1928. URL: http://www.ingentaconnect.com/content/dav/aaua/1995/00000081/00000004/art00009.

Madhu, N. (2009). "Note on measures for spectral flatness". In: *Electronics Letters* 45.23, pp. 1195–1196.

Makhoul, J. I. and J. J. Wolf (1972). *Linear prediction and the spectral analysis of speech*. Tech. rep. 2304. Bolt Beranek and Newman Inc.

Marschner, S. (2012). *Radiometry*. https://www.cs.cornell.edu/courses /cs6630/2015fa/notes /02radiom.pdf. Cornell University. URL: http://www.astrohandbook.com/.

Martellotta, F. (2010). "The just noticeable difference of center time and clarity index in large reverberant spaces." In: *The Journal of the Acoustical Society of America* 128.2, pp. 654–663. ISSN: 1520-8524. URL: http://view.ncbi.nlm.nih.gov/pubmed/20707435.

McCluney, R. (2014). "Introduction to Radiometry and Photometry". In: Second. Artech House, pp. 7–20. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/B00QH2UOXI.

Menzer, F. (2010). "Binaural Reverberation Using Two Parallel Feedback Delay Networks". In: *Audio Engineering Society 40th International Conference: Spatial Audio: Sense the Sound of Space*. Tokyo. URL: https://www.google.com.sg/search?client=safari&rls=en&q=Binaural+Reverberation+Using+

Two+Parallel+Feedback+Delay+Networks&ie=UTF-8&oe=UTF-8&gfe_rd=cr&ei=pChhV9q0Fq6Q8Qe73LPwCw.

Menzer, F. (2012). "Efficient Binaural Audio Rendering Using Independent Early and Diffuse Paths". In: *132 Audio Engineering Society Convention*. 8584. URL: http://www.aes.org/e-lib/browse.cfm?elib=16222.

Menzer, F. and C. Faller (2009). "Binaural Reverberation Using a Modified Jot Reverberator with Frequency-Dependent Interaural Coherence Matching". In: *126 Audio Engineering Society*. 7765. URL: http://www.aes.org/e-lib/browse.cfm?elib=14961.

Miller, George A. (1947). "Sensitivity to Changes in the Intensity of White Noise and Its Relation to Masking and Loudness". In: *The Journal of the Acoustical Society of America* 19.4, pp. 609–619. ISSN: 0001-4966. DOI: 10.1121/1.1916528. URL: http://dx.doi.org/10.1121/1.1916528.

Miller, George A. and Walter G. Taylor (1948). "The Perception of Repeated Bursts of Noise". In: *The Journal of the Acoustical Society of America* 20.2, pp. 171–182. ISSN: 0001-4966. DOI: 10.1121/1.1906360. URL: http://dx.doi.org/10.1121/1.1906360.

Murphy, Damian et al. (2008). "Hybrid Room Impulse Response Synthesis in Digital Waveguide Mesh Based Room Acoustics Simulation". In: *Proc. of the 11 th Int. Conference on Digital Audio Effects (DAFx-08)*. Espoo, Finland.

Nicodemus, F. E. et al. (1977). "Geometrical Considerations and Nomenclature for Reflectance". In: ed. by Lawrence B. Wolff, Steven A. Shafer, and Glenn Healey. Washington, D.C.: Institute for Basic Standards National Bureau of Standards. Chap. II. A. Pp. 3–6. ISBN: 0-86720-294-7. URL: http://portal.acm.org/citation.cfm?id=136929.

Pelzer, S. et al. (2014). "Interactive Real-Time Simulation and Auralization for Modifiable Rooms". In: *Building Acoustics* 21.1. http://dx.doi.org/10.1260/1351-010X.21.1.65, pp. 65–73. URL: http://bua.sagepub.com/content/21/1/65.abstract.

Pelzer, Sönke and Michael Vorländer (2010). "Frequency and Time-dependent Geometry for Real-time Auralizations". In: *Proceedings of 20th International Congress on Acoustics (ICA)*. Sydney, Australia.

Pollack, Irwin (1969). "Periodicity Pitch for Interrupted White Noise—Fact or Artifact?" In: *The Journal of the Acoustical Society of America* 45.1, pp. 237–238. ISSN: 0001-4966. DOI: 10.1121/1.1911363. URL: http://dx.doi.org/10.1121/1.1911363.

Raghuvanshi, N., R. Narain, and M. C. Lin (2009). "Efficient and Accurate Sound Propagation Using Adaptive Rectangular Decomposition". In: *IEEE Transactions on Visualization and Computer Graphics* 15.5, pp. 789–801. URL: https://ieeexplore.ieee.org/document/5165582/.

Raghuvanshi, Nikunj et al. (2010). "Precomputed Wave Simulation for Real-time Sound Propagation of Dynamic Sources in Complex Scenes". In: *ACM Transactions on Graphics* 29.4. https://dx.doi.org/10.1145/1778765.1778805. ISSN: 0730-0301. DOI: 10.1145/1778765.1778805. URL: http://dx.doi.org/10.1145/1778765.1778805.

Rizzi, S. A. (2013). *NASA Technical Reports Server (NTRS) - An Overview of Virtual Acoustic Simulation of Aircraft Flyover Noise*. Tech. rep. Pisa, Italy. URL: http://ntrs.nasa.gov/search.jsp?R=20140000602.

Rocchesso, D. (1997). *Maximally diffusive yet efficient feedback delay networks for artificial reverberation*.

Rocchesso; D. and J. O. Smith (1997). "Circulant and elliptic feedback delay networks for artificial reverberation". In: *IEEE Transactions on Speech and Audio Processing* 5.1, pp. 51–63.

Rochesso, D. (2000). "Fractionally addressed delay lines". In: *IEEE Transactions on Speech and Audio Processing* 8.6, pp. 717–727.

Roth, S. D. (1982). "Ray Casting for Modeling Solids". In: *Computer Graphics and Image Processing* 18.2. https://doi.org/10.1016/0146-664X(82)90169-1.

Sarti, A. and S. Tubaro (2001). "Low-Cost Geometry-Based Acoustic Rendering". In: *Proc. of the COST G-6 Conference on Digital Audio Effects (DAFX-01)*. Limerick, Ireland. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.6992.

Savioja, L., T. Lokki, and J. Huopaniemi (2002). "Auralization applying the parametric room acoustic modeling technique - The DIVA auralization system". In: *Proc. International Conference on Auditory Display 2002*. Kyoto, japan.

Savioja, L. and U. P. Svensson (2015). "Overview of geometrical room acoustic modeling techniques". In: *The Journal of the Acoustical Society of America* 138.708.

Savioja, L. et al. (1999). "Creating interactive virtual acoustic environments". In: *Journal of the Audio Engineering Society* 47.9, pp. 675–705. URL: http://lib.tkk.fi/Diss/2002/isbn9512261588/article1.pdf.

Schröder, D. (2011). "Physically based real-time auralization of interactive virtual environments". PhD thesis. Fakultät für Elektrotechnik und Informationstechnik der Rheinisch-Westfälischen Technischen Hochschule Aachen. URL: https://www.google.com.sg/search?client=safari&rls=en&q=raven+pdf+acoustic+paper&ie=UTF-8&oe=UTF-8&gfe_rd=cr&ei=vHegWI2xOIq2vATSuKmACQ#q=interactive+real+time+dirk+schroder.

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

Schröder, Dirk, Philipp Dross, and Michael Vorländer (2007). "A Fast Reverberation Estimator for Virtual Environments". In: *30th International Conference: Intelligent Audio Environments*. 13.

Schröder, Dirk and Tobias Lentz (2006). "Real-Time Processing of Image Sources Using Binary Space Partitioning". In: *Journal of the Audio Engineering Society* 54.7/8, pp. 604–619.

Schroeder, M. et al. (2007). "Springer Handbook of Acoustics". In: 1st. Springer Publishing Company, Incorporated. Chap. 11.1.4, pp. 394–395. ISBN: 0387304460, 9780387304465. URL: http://portal.acm.org/citation.cfm?id=1535484.

Schroeder, M. R. (1960). ""Colorless" Artificial Reverberation". In: *The Journal of the Acoustical Society of America* 32, pp. 1520+. DOI: 10.1121/1.1936376. URL: http://dx.doi.org/10.1121/1.1936376.

— (1962). "Natural Sounding Artificial Reverberation". In: *Journal of the Audio Engineering Society* 10.3, pp. 219–223.

Schroeder, Manfred R. and Benjamin F. Logan (1961). "'Colorless' Artificial Reverberation". In: *Journal of the Audio Engineering Society* 9.3, pp. 192–197.

Searls, Donald T. (1966). "An Estimator for a Population Mean Which Reduces the Effect of Large True Observations". In: *Journal of the American Statistical Association* 61.316, pp. 1200–1204. DOI: 10.1080/01621459.1966.10482204. URL: http://dx.doi.org/10.1080/01621459.1966.10482204.

Sena, E. D. et al. (2015). "Efficient Synthesis of Room Acoustics via Scattering Delay Networks". In: *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 23.9, pp. 1478–1492. ISSN: 2329-9290. DOI: 10.1109/taslp.2015.2438547. URL: http://dx.doi.org/10.1109/taslp.2015.2438547.

Shaw, E. A. and R. Teranishi (1968). "Sound pressure generated in an external-ear replica and real human ears by a nearby point source." In: *The Journal of the Acoustical Society of America* 44.1, pp. 240–249. ISSN: 0001-4966. URL: http://view.ncbi.nlm.nih.gov/pubmed/5659838.

Siltanen, S., T. Lokki, and L. Savioja (2010). "Rays or Waves? Understanding the Strengths and Weaknesses of Computational Room Acoustics Modeling Techniques". In: *Proc. Int. Symposium on Room Acoustics (ISRA)*. Melbourne, Australia.

Siltanen, S. et al. (2007). "The room acoustic rendering equation." In: *The Journal of the Acoustical Society of America* 122.3. https://doi.org/10.1121/1.2766781, pp. 1624–1635. ISSN: 1520-8524. DOI: 10.1121/1.2766781. URL: http://dx.doi.org/10.1121/1.2766781.

Smith, J. O. (2010). *Physical audio signal processing for virtual musical instruments and audio*. W3K Publishing. ISBN: 0974560723.

South, S. C., T. F. Oltmanns, and E. Turkheimer (2005). "Interpersonal Perception and Pathological Personality Features: Consistency Across Peer Groups". In: *Journal of Personality* 73.3, pp. 675–692. ISSN: 0022-3506. DOI: 10.1111/j.1467-6494.2005.00325.x. URL: http://dx.doi.org/10.1111/j.1467-6494.2005.00325.x.

Stelmachowicz, P. G. et al. (1989). "Normative thresholds in the 8- to 20-kHz range as a function of age". In: *Journal of the Acoustical Society of America* 86.4, pp. 1384–1391.

Stevens, S. S. and J. Volkmann (1937). "A Scale for the Measurement of the Psychological Magnitude Pitch". In: *The Journal of the Acoustical Society of America* 8.3, pp. 185–190.

Tenenbaum, R. A. et al. (2007). "Hybrid Method for Numerical Simulation of Room Acoustics: Part 2 – Validation of the Computational Code RAIOS 3". In: *Journal of the Brazilian Society of Mechanical Sciences and Engineering* 29.2, pp. 222–231. URL: https://www.semanticscholar.org/paper/Hybrid-Method-for-Numerical-Simulation-of-Room-Aco-Tenenbaum-Camilo/a3bc9f6de9961e369019b1f

Välimäki, V. et al. (2012). "Fifty Years of Artificial Reverberation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.5. https://doi.org/10.1109/TASL.2012.2189567, pp. 1421–1448. ISSN: 1558-7916. DOI: 10.1109/tasl.2012.2189567. URL: http://dx.doi.org/10.1109/tasl.2012.2189567.

Välimäki, Vesa et al. (2016). "More Than 50 Years of Artificial Reverberation". In: *60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*. K-1.

Vigeant, M. C. et al. (2015). "The effects of different test methods on the just noticeable difference of clarity index for musica)". In: *The Journal of the Acoustical Society of America* 138.1. http://dx.doi.org/10.1121/1.4922955, pp. 476–491. ISSN: 0001-4966. DOI: 10.1121/1.4922955. URL: http://dx.doi.org/10.1121/1.4922955.

Vorländer, Michael (1989). "Simulation of the transient and steady state sound propagation in rooms using a new combined ray tracing image source algorithm". In: *The Journal of the Acoustical Society of America* 86.1. https://doi.org/10.1121/1.398336, pp. 172+.

Wallach, H., E. B. Newman, and M. R. Rosenzweig (1949). "The precedence effect in sound localization". In: *The American Journal of Psychology* 62, pp. 315–336.

Wang, L. M., J. Rathsam, and S. Ryherd (2004). "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*. Awaji Island, Japan. URL: https://nebraska.pure.elsevier.com/en/publications/the-influence-of-absorption-factors-on-the-sensitivity-of-a-virtu.

Wendt, T., S. Van de Par, and S. Ewert (2014a). "A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation". In: *Journal of the Audio Engineering Society* 62.11. https://doi.org/10.17743/jaes.2014.0042, pp. 748–766. ISSN: 15494950. DOI: 10.17743/jaes.2014.0042. URL: http://dx.doi.org/10.17743/jaes.2014.0042.

— (2014b). "Perceptual and Room Acoustical Evaluation of a Computational Efficient Binaural Room Impulse Response Simulation Method". In: *Proc. of the EAA Joint Symposium on Auralization and Ambisonics*. https://doi.org/10.14279/depositonce-15. Berlin, Germany. URL: https://depositonce.tu-berlin.de/handle/11303/172.

Wicke, Roger W. and Adrianus J. M. Houtsma (1975). "Musical pitch of interrupted white noise". In: *The Journal of the Acoustical Society of America* 58.S1, S83. ISSN: 0001-4966. DOI: 10.1121/1.2002346. URL: http://dx.doi.org/10.1121/1.2002346.

Wilson, E. B. and M. M. Hilferty (1931). "The distribution of chi-square". In: *Proceedings of the National Academy of Sciences of the United States of America* 17, pp. 684–688.

Woirgard, M. et al. (2012). *Cologne University of Applied Sciences - Anechoic Recordings*.

Zwicker, E. (1961). "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)". In: *The Journal of the Acoustical Society of America* 33.2, pp. 248+.