Spatial Sound Rendering Using Measured Room Impulse Responses

by

Yan Li
B.Eng, Northwestern Polytechnical University, 1996
M.Eng, Dalian University of Technology, 1999

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Applied Science

in the Department of Electrical and Computer Engineering

Spatial Sound Rendering Using Measured Room Impulse Responses

by

Yan Li
B.Eng, Northwestern Polytechnical University, 1996
M.Eng, Dalian University of Technology, 1999

Supervisory Committee

Dr. Peter F. Driessen, Co-Supervisor
(Department of Electrical and Computer Engineering)

Dr. George Tzanetakis, Co-Supervisor
(Department of Computer Science)

Dr. Wu-Sheng Lu, Departmental Member
(Department of Electrical and Computer Engineering)

Supervisory Committee

Dr. Peter F. Driessen, Co-Supervisor
(Department of Electrical and Computer Engineering)

Dr. George Tzanetakis, Co-Supervisor
(Department of Computer Science)

Dr. Wu-Sheng Lu, Departmental Member
(Department of Electrical and Computer Engineering)

ABSTRACT

This thesis presents a spatial sound rendering system for the use in immersive virtual environments. Spatial sound rendering aims at artificially reproducing the acoustics of a space. It has many applications such as music production, movies, electronic gaming and teleconferencing. Conventionally, spatial sound rendering is implemented by digital signal processing algorithms derived from perceptual models or simplified physical models. While being flexible and/or efficient, these models are not able to capture the acoustical impression of a space faithfully. On the other side, convolving the sound sources with properly measured impulse responses produces the highest possible fidelity, but it is not practically useful for many applications because one impulse response corresponds to one source/listener configuration so that the sources or the listeners can not be relocated.

In this thesis, techniques for measuring multichannel room impulse responses (MMRIR) are reviewed. Then, methods for analyzing measured MMRIR and rendering virtual acoustical environment based on such analysis are presented and evaluated. The analysis can be performed off-line. During this stage, a set of filters that represent the characteristics of the air and walls inside the acoustic space are obtained. Based on the assumption that the MMRIR acquired at one "good" position in the target space can be used to simulate the late reverb at other positions in the same space, appropriate segments that can be used as reverb tails are extracted from the measured MMRIR. The rendering system first constructs an early reflection model based on the positions of the listener-source pair and the filters derived, then combines with the late

reverb segments to form a complete listener-source-room acoustical model that can be used to synthesize high quality multi-channel audio for arbitrary listener-source positions. Another merit of the proposed framework is that it is scalable. At the expense of slightly degraded rendering quality, the computational complexity can be greatly reduced. This makes this framework suitable for a wide range of applications that have different quality and complexity requirements.

The proposed framework has been evaluated by formal listening tests. These tests have proven the effectiveness in preserving the spatial quality while positioning the listener-source pair accurately, as well as justified the key assumptions made by the proposed system.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

It is a pleasure to thank the many people who made this thesis possible.
It is difficult to overstate my gratitude to my supervisor, Dr. Peter Driessen, whose expertise, understanding, and patience, added considerably to my graduate experience. He has made available his support all the ways I can ask for.
I would like to express my gratitude to my co-supervisor, Dr. George Tzanetakis, for his enthusiasm and encouragement. His inspiration and ideas made the thesis work a more enjoyable journey. I would like to thank Dr. Wusheng Lu, for the inspirational and fun courses and ideas he offered for the course projects.
I would like to thank Kirk McNally, who offered valuable feedback and advice in many areas touched by this thesis. I would also like to thank all the participants in the subjective evaluation and staff at Banff center for access to the recital halls and equipment to do the recording.
I owe my deepest gratitude to my family. My wife, Li, has been very supportive throughout my lengthy graduate program, as she has always been for the past 12 years. This thesis would not have been possible without her love, care and patience. I wish to thank my parents who have been a constant source of support throughout my life. Last, I would like to thank my son, Samuel, who brings endless joy to every single day.

To my family.

Chapter 1

Introduction

In recent years there has been significant interest in the synthesis of immersive virtual environments. Applications for this technology include entertainment, communication, remote control, and simulation. It is essential that these simulations include a realistic recreation of the intended auditory scene. The goal of spatial sound rendering is to create a virtual auditory environment that is indistinguishable from a real auditory environment. In this thesis, I present a spatial sound rendering system based on the multi-channel measured room impulse response (MMRIR) with the goal of creating the acoustical impression of a specific venue using a multichannel speaker system for arbitrary sources and listener positions.

1.1 What Is Spatial Sound Rendering?

Spatial sound rendering, or auralization in some literature [1] [2] [3], has been defined as "*the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space*" [1]. However, the problem here in discussion differs slightly from the normal definition of *Auralization*. Firstly, perceptual or psycho-acoustical modeling is allowed because the goal is to "deceive" the human brain about the acoustic impression rather than to find an accurate physical or mathematical approximation. As discussed in later chapters, such modeling methods do not satisfy the requirement of accurate reproduction of an acoustic space. Secondly *Auralization* is often referred to as a CAD tool for simulating room acoustics of an architectural subject in order to study its acoustical behavior before being

built. To take these differences into account, the term *Spatial Sound Rendering* is used instead throughout this thesis.

The goal of spatial sound rendering is to create a virtual auditory environment that is indistinguishable from a real auditory environment and to accurately position the sound sources in this virtual auditory environment. From a digital signal processing point of view, the goal is to process and convert a set of anechoic mono audio signals[1] to multichannel signals to be played on an appropriate speaker system or headphones, in such a way that the perceived sound sources and acoustical impression are indistinguishable from what the listener would perceive in the real environment.

1.1.1 Requirements and Challenges

There are five conditions necessary to achieve realism in a sound reproducing system [4]:

1. The frequency range of the reproduced sound should be sufficient to retain all the audible components in the source sound, and the sound spectrum of the reproduced sound should be identical to that of the source.

2. The reproduced sound should be free of distortion and noise.

3. The reproduced sound should have loudness and dynamic range comparable to the original sound.

4. The spatial sound pattern of the original sound should be reproduced.

5. The reverberation characteristics (in space and time) of the original sound should be preserved in the reproduced sound.

Modern sound reproducing equipment can satisfactorily achieve the first three conditions [5]. It is the latter two conditions that pose challenges to the design of audio systems, particularly systems that are intended for extremely realistic spatial reproduction of audio.

The **spatial pattern** depends on the positions of the sound sources and the listeners, the radiation patterns of the sound sources and the directional patterns of the receiver, be it listener or microphone. A system that is capable of reproducing the

[1] Audio signals that are recorded using close microphones in anechonic rooms can be considered anechonic.

spatial pattern is often referred to as a 3D audio system. In order to create convincing immersive virtual environments, the perceived position of a sound source created by the 3D audio system must match the spatial location of the associated object in the virtual environment. Designing a realistic 3D audio system can be very challenging. Its main difficulties include:

1. the system must be able to map sound sources at arbitrary virtual positions to a limited number of speakers;

2. the overall acoustic impression must match that of the intended acoustic space;

3. it needs to be efficient in order to handle large amount of sound sources simultaneously;

4. in an interactive application, it needs to handle movement of the sound sources or the listener smoothly.

The spatial pattern in a 3D audio system is primarily determined by the direct arrival of the sound. It is also affected by the early reflections (see below) but to a much lesser degree.

The **reverberation** refers to the prolongation or persistence of sound within an enclosure as sound waves reflect off hard surfaces (bare walls, ceilings, windows and floors) in the room [6]. A reverberation, or reverb, is created when a sound is produced in an enclosed space causing a large number of echoes to build up and then slowly decay as the sound is absorbed by the walls and air [7]. Its characteristics is primarily determined by the properties of the acoustic space such as room geometry and surface material, among others. The presence of reverberation is clearly preferred for most sounds, particularly music. Music without reverberation sounds dry and lifeless. On the other hand, too much reverberation, or inappropriate reverberation, can cause a fine musical performance to sound muddy and unintelligible [8].

Reverberation sometimes can be divided into two segments, early reflections and late reverberation. Early reflections refer to the reflections that occur soon after the initial sound. They are generally a set of well defined and directional reflections that are directly related to the shape and size of the room, as well as the position of the source and listener in the room. Instead of being perceived as separate sound events, early reflections modify the perception of the sound, changing the loudness, timbre, and most importantly, the spatial characteristics of the sound [8]. After the early

reflections, the rate of the arriving reflections increases greatly. These reflections are more random and difficult to relate to the physical characteristics of the room. This is called the diffuse reverberation, or the late reflections. It is believed that the diffuse reverberation is the primary factor establishing the perception of the 'size' of an acoustic space[9].

The spatial pattern and the reverberation are closely related and affect the perception of the sounds and space in a combined fashion. From a digital signal processing perspective, it is often convenient to re-organize them into three segments of a room impulse response as depicted in Fig. 1.1, namely direct sound, early reflections and late reverberation. Each segment corresponds to a distinct type of path along which the sound travels.

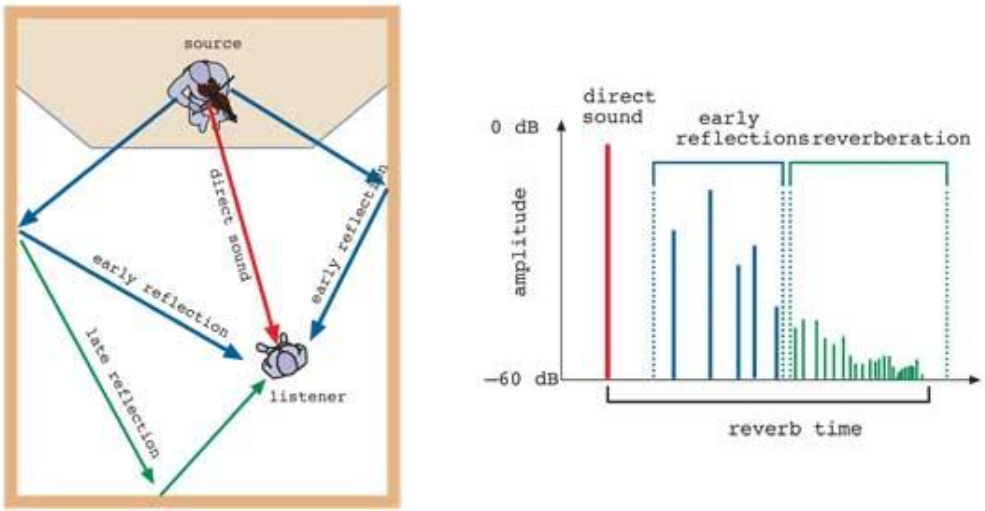


Figure 1.1: A typical room impulse response and sound travel paths[10]

As to be discussed in the latter chapters of this thesis, these two challenges can be attacked individually, with somewhat satisfactory results. However, when combined together, they create much more difficulties that can not be solved by the methods that are designed to solve each one separately. This thesis concerns the design, implementation, and evaluation of an spatial sound rendering system that attempts to simultaneously address these challenges, namely, reproducing the reverberation characteristics and the spatial pattern faithfully.

1.1.2 Psychoacoustics of Spatial Hearing

The ultimate goal of spatial sound rendering is to create convincing acoustical events and cues for the human ears and brains, using a set the limited speakers or headphones. Therefore, the designer of such system must understand what influences the human auditory system, more specifically, how the human auditory system locates the sound sources and how it perceives the surround environment. Several important cues that affect the spatial perception are reviewed briefly.

Azimuth

According to the Duplex Theory [11] of spatial sound perception [2], there are two primary cues for azimuth – Inter-aural Time Difference (ITD) and Inter-aural Intensity Difference (IID) that the hearing system uses for estimating the apparent horizontal direction of a sound source. When the sound is not received directly from the front or from the back, it travels a longer distance to one ear than the other, and therefore does not arrive at both ears at the same time, but with a time difference. For instance, as depicted in 1.2, given the angle of arrival of θ, the difference in the traveling distance $\Delta d = d_2 - d_1$ results in a difference in arrival time $\Delta t = \frac{\Delta d}{c}$ where c is the speed of sound. IID, on the other hand, is the intensity difference between the two ears that is caused by the head-shadow effect, where the far ear is in the sound shadow of the head. The Duplex Theory also states that the IID and the ITD are complementary. At low frequencies (below about 1.5kHz), there is little IID information, but the ITD shifts the waveform a fraction of a cycle, which is easily detected. At high frequencies (above about 1.5 kHz), there is ambiguity in the ITD, since there are several cycles of shift, but the IID resolves this directional ambiguity.

Elevation

Judging the elevation (the angle between the source and the horizontal plane through both ears) is more complex. The determining factor of elevation perception is the "pinna notch". Our outer ear, or pinna, acts as an acoustical resonant filter [11]. Depending on the incoming elevation, certain frequencies may constructively or destructively interfere within its resonant cavities, adding peaks and notches to its transfer function. Given a wide-band sound source, the human auditory system picks up these subtle frequency dependent cues and use them to determine the elevation.

[2] There is a Duplex Theory of pitch.

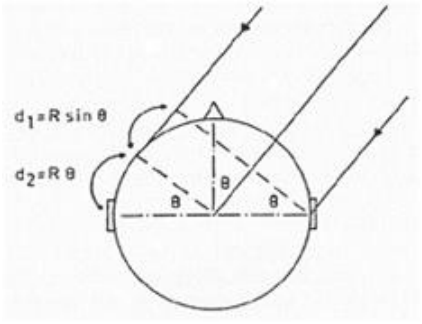


Figure 1.2: Inter-aural Time Difference (ITD)

There are also other cues and methods that supplement the "pinna notch" cue, such as shoulder notches and head tilting. Unlike the azimuth cues which are to a large extend listener independent, elevation cues are more specific to the particular listener. The spectra cues that influence the spatial hearing can be empirically measured as Finite Impulse Response (FIR) banks, which are known as the Head Related Transfer Functions (HRTFs) [12].

Distance

The perception of distance is even more complex. It relies on more cues and these cues always come with ambiguities and tend to interact with each other. There are several frequently mentioned cues that are used to estimate the distance of a sound source.

1. As a constant-energy source approaches a listener, the **loudness** will increase. Although it provides useful information, we can not use it alone to determine the distance because the loudness depends not only on the distance, but also on the energy emitted by the source.

2. **Motion parallax** is a dynamic cue that refers to the change in azimuth of the sound source resulting from translation of a source or listener. It only marginally improves distance perception accuracy [13]. For sources that are very close, a small shift causes a large change in azimuth, while for sources that are distant there is essentially no azimuth change [11].

3. **Excessive IID** refers to the increased IID when a sound source comes very close to the head. In general, sounds that are heard in only one ear are threatening

and are uncomfortable to listen to.

4. **Ratio of direct to reverberant sound** is another cue to distance in rooms with reflective walls. In general, the greater the distance of the source, the greater is the proportion of reflected sound. The direct-to-reverberant ratio is a major cue for distance [11]. A detailed study on this topic can be found in [14].

5. Large changes in distance tend to change the **spectral balance** of the sound reaching the ears [15]. This is due to the fact that the air absorbs more high-frequency energy than low-frequency energy. So given a wide-band source, it sounds brighter when it is closer to the listener. Depending on the geometry of the space, diffraction may also contribute to the loss of the high-frequency energy because higher frequencies do not diffract as much as lower frequencies.

In any natural listening environment, the auditory system will tend to combine all these cues to estimate distance. The loudness and direct-to-reverberant ratio are the two more salient cues [15]. For a sound source very close to the listener, the auditory system can estimate the distance fairly accurately. It is still the case, however, that for distance greater than a meter, we seem to consistently underestimate the true distance when relying on acoustic information alone [15].

Environment

Reverberation which includes the early reflections and late reverberation discussed earlier, provides crucial information of the space such as the room dimensions and the surface materials. Even outdoors, a significant amount of energy is reflected by the ground and by surrounding structures and vegetation. In a properly designed real or virtual acoustic space, the reverberation adds fullness and life to the acoustic events, and sometimes even masks imperfection in, for example, a vocal performance. All these merits are offered without degrading our ability to locate the sound sources [3]. This is largely due to the fact that we localize the sound source based on the signal that reaches our ears first, known as the "precedence effect"[11].

[3]Human listeners can robustly localize sound sources in a moderately reverberant environment [16]

Summary

In general, humans are best at judging sound source azimuth, then elevation, and worst at judging distance. Inspired by Table.3 in [17], the influence of direct sound, early reflections and late reverb (as mentioned in the previous section) on different levels of perception is summarized in Table 1.1 [4].

	Direct sound	Early reflections	Late reverberation
Horizontal direction	• • •	•	
Elevation	○ ○ ○	○	
Near-head distance	○ ○ ○		
Distance, spatial depth	•	• • •	•
Spatial impression		• •	• • •
Envelopment			• •
Sound color	•	• •	• • •

Table 1.1: Influence of Reverberation and Spatial Pattern on Different Levels of Perception

The method presented in this thesis attempts to reproduce these directional, distance and environmental cues accurately to create a convincing virtual audio space.

1.1.3 Headphones or Loudspeakers

There are two options for reproducing electronic sounds, loudspeakers and headphones. Each of them has unique advantages and disadvantages when used for reproducing spatial sounds. The following section gives a brief review and comparison between loudspeaker systems and headphones.

A headphone provides an isolated listening environment that excludes noise and interference from the acoustic events and acoustical characteristics within which the listener is located. It also has a fixed relative position to the listener's ear, which allows the listener to move around without re-adjusting the headphones. However, headphones also have several shortcoming that are very difficult to overcome. Due to the size limitation, headphones use tiny transducers and place the transducer very close to the ear. Tiny transducers generally are not able reproduce the full audible frequency range, especially the low frequencies. Having the transducers right next to ears means most of the spatial resolution is lost [18] [19]. Stereo headphones

[4] In this table, "○" indicates that the corresponding aspect is not addressed in this thesis.

(one transducer at each ear) are not able to reproduce spatial cues beyond the 1-dimensional space which consists of the line from left ear to right ear. In order to reconstruct a 3D sound field to a certain degree, sophisticated signal processing such as Head Related Transfer Function (HRTF) filtering is required to simulate the path from the sound sources to the ear. However, to accurately simulate the path demands the signal processing system to be adapted to each individual listener [20] [21] and each pair of headphones. Multi-driver headphones have been available for some time, for example Razer Megalodon 7.1 and LogicTech G35[22]. While they seem to be able to reconstruct portion of the spatial resolution [22], these "surround headphones" also suffer from the same principle limitations.

Stereo loudspeakers system started to prevail in 1950's. With stereo panning in conventional stereo systems, sounds can be panned to locations between the two loudspeakers, creating virtual or phantom images of the sound where there is no loudspeaker. However, conventional stereo systems generally cannot position sounds to the sides or rear of the listener, nor above or below the listener. With the intent to extend the positionable area, [5] proposed a 3-D audio system that synthesizes and playbacks binaural signals to conventional stereo speakers. Such an approach requires tracking the listener's position and applying crosstalk cancellation to invert the transmission paths that exist from the speakers to the listener. These requirements impose practical difficulties to implement and may downgrade audio quality dramatically. Since their introduction and standardization in 1990s, multichannel surround sound systems such as 5.1 and 7.1 systems have quickly become mainstream configuration for home theater environments. According to online user poll [23], such systems make up the majority of home theater and gaming audio setups. The advantages of surround speaker systems are obvious. They provide physical speakers all around the listeners, including at the back so that the positionable area is naturally extended to all four quadrants. Unlike the 3D system based on stereo systems, surround speakers do not used binaural signals and thus eliminate the need of crosstalk cancellation. Because the majority of the surround speakers systems are set up in home theater or gaming environments that have video content on screen playing at the direction of the center speaker as in Fig. 1.3, the listeners' positions become relative to the screen (and in turn the speakers) and need not to be tracked and adapted to as long as the listeners remain within the perimeter of the speaker systems.

Based the above observation, this thesis focuses on building a 3D system based on surround speakers, more specifically, 5.1 configurations based on an ITU standard [24]

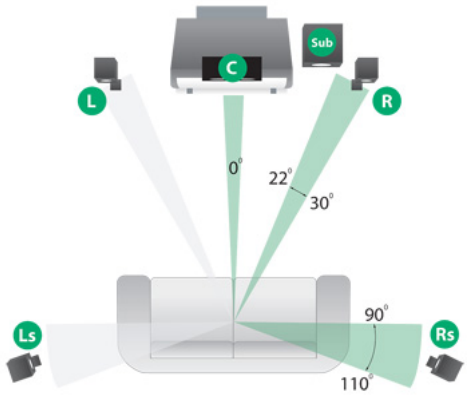


Figure 1.3: Standard Dolby Digital 5.1 Setup For Home Theatre

or alternatively 5 equally spaced configuration. However, such a system can be easily adapted to any configuration and can also be extended to including Head Related Transfer Functions (HRTFs) so that it can be used with the headphones.

1.2 Applications

Spatial sound rendering has many applications in the areas of virtual reality and tele-presence. Spatial sound could help increase the sense of presence in virtual environments by relaying information about the environment and the objects within it. Such environmental awareness could be very beneficial in increasing the user's orientation in virtual environments [25]. Specifically, these applications include, but are not limited to, electronic games, movie/music production, teleconferencing, networked music performance, audio-based navigation interfaces for the blind, and architectural simulation. Different application areas have vastly different complexity and quality requirements. For example, music/movie production requires the highest possible quality and generally has more available computational power and no real-time constrains, while on the other side, for computer and video gaming applications, the rendering system must handle very complex scenarios in real-time that contain thousands of dynamic sound sources and fast changing environments with limited computing resources. This thesis aims at providing a unified framework that is able to provide high quality as well as being flexible and efficient enough to process complex scenarios.

1.3 Contribution and Organization of the Thesis

The main contributions of the thesis include the proposition and evaluation of a hybrid spatial sound rendering solution that combines convolution based approaches and ray-tracing based approaches, as well the design of a real-time system that realizes this hybrid solution. More specifically, in this thesis I present a new spatial sound rendering system based on the MMRIR with the goal of creating the acoustic impression of a specific venue using a multichannel speaker system. To achieve this goal, a hybrid method is proposed that models only the direct sound and early reflections individually using the image-source method and synthesizes the late reverberation using a set of filters derived from the MMRIR. Unlike other solutions, this system is built exclusively on the MMRIR - the true reflection of the acoustic characteristics of the target venue. The effectiveness of this system is proven by formal subjective evaluations and efficiency proven by a real-time implementation.

The thesis is organized as follows. Chapter 2 provides a review of existing techniques that try to achieve our goal. The measurement and analysis the MMRIR is described in Chapter 3, followed by how the room acoustic model is constructed using the analysis results in Chapter 4. In Chapter 5, I elaborate the system design and implementation, followed by evaluation and discussions.

Chapter 2

State of the Art

The research field related to spatial sound rendering is multidisciplinary, thus the design and implementation of the system requires understanding and knowledge of room acoustics, digital signal processing, and psychoacoustics [2]. In the past two decades, a significant amount of research has been carried out in the areas related to spatial sound rendering. However, the focus of this research went to two extremes. Traditionally research into spatial sound has focused upon high quality renderings of the spatial environment. Spatial rendering has primarily been based upon geometrical properties of environments, physical properties of objects, and source characteristics, e.g [26]. This approach, whilst very accurate, requires powerful processing resources and is very difficult to achieve in real-time applications [27]. At the other end, a number of real-time rendering systems have been proposed, mainly for the purpose of electronic gaming [28] [29]. These real-time rendering systems are often built on (overly) simplified perceptual or physical model and faintly resemble the physical reality. For example, earlier artificial reverberators [30] used parallel comb filters and cascaded all-pass filters to synthesize reverberation, Creative's EAX uses a Feedback Delay Network (FDN) which can be viewed as a network of multi-channel comb filters [28] and A3D only simulates the first few reflections [29]. In the context of Virtual Aural Reality, a number of projects targeted at rendering "good" quality at reasonable complexity so as to be implemented in real-time have made advances in different areas, for example, DIVA [31] utilizes a parametric RIR rendering method. Tsingos et al [32] proposed a real-time 3D audio rendering pipeline that is capable of rendering complex virtual scenes containing hundreds of moving sound sources. However, these systems are still built on, or partially built on, the imaginary or theoretical models that may not always reflect the physical reality. Also having to control a parametric

model often troubles users who do not fully understand the impact of each of the parameters[33].

These two categories offer distinct advantages and suffer from unique disadvantages so that given the current state of technology, no one can replace the other. This is why the MPEG-4 standard supports both perceptual and physical (geometrical) room models [34]. Having to support both models imposes great challenges to its implementation, especially when the requirements are high quality, precise synchronization with other media and acceptable latency to user interaction, as it often happens in standardized contexts for media integration [35]. The rendering model proposed in this thesis aims to provide one solution that is capable of producing highest quality as well as being scalable and flexible to fit in a wide range of applications.

2.1 Perceptual Approaches

The spatial acoustical impression is created by the interaction between the sound waves, the environment and the listeners. Because it can be extremely difficult to accurately model and recreate the true physics of such interaction, researchers have developed alternatives that seek to reproduce only the perceptually salient characteristics of the room acoustics.

Perceptual based approaches share these common advantages

1. The rendering algorithms can be implemented efficiently, for exampling, using amplitude-only panning for direct sounds and using infinite impulse response (IIR) filters for late reverberation;

2. The rendering algorithms often provide real-time control of all the perceptually relevant parameters. The parameters do not need to be correlated as they often are in real situation;

3. Ideally, only one set of modeling and rendering algorithms is required to simulate many possible scenarios.

Over the past few decades, various perceptual based algorithms have been developed to cover all aspects of spatial sound rendering [30] [28] [36] [37] [32]. However, our knowledge on spatial sound perception is very limited and rudimentary, therefore there is a great deal of disagreement as to what the perceivable attributes of reverberation are and how to measure these attributes. Beyond that, it is difficult to design

digital filters to reproduce these attributes [8]. Controlling these attributes in a physically meaningful way may require significant efforts, for example, mulit-dimensional nonlinear optimization techniques have been used to automatically tune a reverberator to match a specific impulse response [33]. Thus, the perceptual based approaches have not gained much interest beyond application areas in which efficiency is at a higher priority than the quality. Their most notable application is video gaming.

2.1.1 Direct Sound and Early Reflections

To be physically accurate, the direct sound and each of the early reflections must be modeled as discrete acoustic events that arrive at each of the receivers (human ears or loudspeakers) at different times and bearing different energy levels.

Vector base amplitude panning (VBAP) method [36] is an amplitude-only panning method that is capable of positioning virtual sources in arbitrary loudspeaker setups. For the setups that all the speakers are on a horizontal plane (referred to as a 2-D setup in [36]), VBAP uses conventional tangent law

$$\begin{cases} g_n = \cos\theta_x \\ g_m = \sin\theta_x \end{cases}$$

to calculate a pair of gains for panning between two adjacent speakers.

In a 3D setup shown in Fig. 2.1, where speakers can be placed above or below the horizontal plane, the tangent law is generalized by vector reformulation so that the panning direction vector $\mathbf{p}$ can be expressed as a linear combination of the gain vector $\mathbf{g} = [g_n \ g_m \ g_k]$ and the three loudspeaker vectors $\mathbf{l}_n$, $\mathbf{l}_m$ and $\mathbf{l}_k$,

$$\mathbf{p}^T = \mathbf{g}\mathbf{L}_{nmk}$$

where $\mathbf{L}_{nmk} = [\mathbf{l}_n \ \mathbf{l}_m \ \mathbf{l}_k]^T$. Therefore the gain vector $\mathbf{g}$ can be solved

$$\mathbf{g} = \mathbf{p}^T\mathbf{L}_{nmk}^{-1}$$

The classic tangent law panning and VBAP have been utilized extensively in the music production and video gaming, but it has very limited use in a complete 3D audio rendering system. Because it does not provide arrival time information, it can not be used for early or late reflections for which timing cues are critical.

For the case of early reflections, it is obvious that these delayed and attenuated

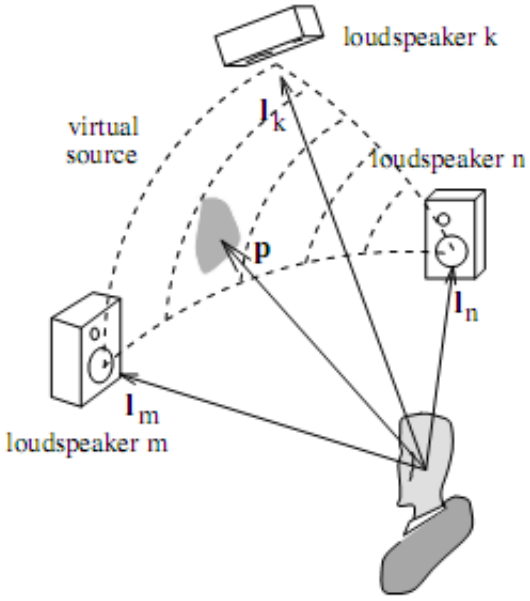


Figure 2.1: VBAP in a 3D setup [36]

copies of the direct sound can be modeled by an FIR filter that has sparsely spaced impulses as in Fig. 1.1. This is a much simplified model derived from the source image model that does not consider the complex interaction between the acoustic waves and the surrounding environment, such as diffraction, diffusion and absorption. However, this model allows for efficient implementation, namely delay-and-sum as shown in Fig 2.2, and has been widely used in many artificial reverberators [30] [38] and 3D audio system [28] [31].

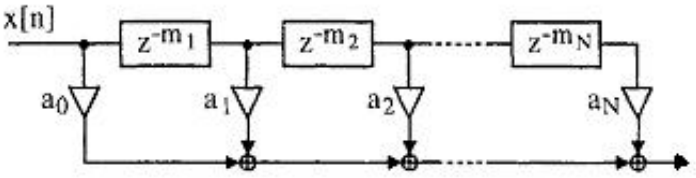


Figure 2.2: Delay-and-sum Implementation of Early Reflections

Technically, though greatly simplified, this FIR reflection model that generates a sparse impulse response is a physically based model. However, the delays m_n and gains a_n in many rendering systems [39] [38] [28] are not derived from physical

model of the room geometry. Instead, they are determined heuristically to create the impression of delayed and attenuated sparse echoes.

One main disadvantage of modeling the early echoes using a sparse FIR filter results in an overly discrete sound quality, particularly with bright impulse inputs [8]. A number of techniques that incorporate physical models have been developed to address this problem. These techniques will be discussed in the next section.

2.1.2 Late Reverberation

Late reverberation is the most difficult and costly component for accurate physical modeling. Therefore the majority of commonly used reverberators are based on perceptual models. Among many criteria of designing practically useful and natural sounding reverberators, the following four are recognized as the most important ones

1. exponentially decaying response with adjustable reverberation characteristics;
2. sufficient reflections density;
3. sufficient modal density or modes (resonances) per herz to avoid coloration, ringing tones, metallic sound;
4. incoherence between different channels (monophonic reverberation does not give a good spatial impression).

In the early days, comb filter structures such as the one shown in Fig 2.3 were used to create reverberation because it can efficiently generate replicas of a direct sound that are time delayed and attenuated. However, Fig 2.4 shows that the replicas produced by comb filter have exactly the same time delay in between. This leads to the sensation of a pitched tone superimposed on the signal [40]. Specifically, its magnitude response is not constant for all frequencies resulting in, a 'coloration' of many musical sounds that are often unpleasant for listening purposes. Another defect is that the output echo density generated by a unit impulse at the input is much lower than that observed in a real room.

A slightly more realistic reverberator can be implemented using an allpass filter structure. All-Pass filters give us the echoes as before, but a smoother frequency response. They have the effect of frequency-dependent delay, smearing the harmonics of the input signal and getting closer to a true reverberation sound.

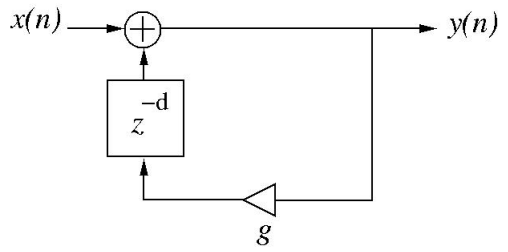


Figure 2.3: IIR Comb Filter Structure

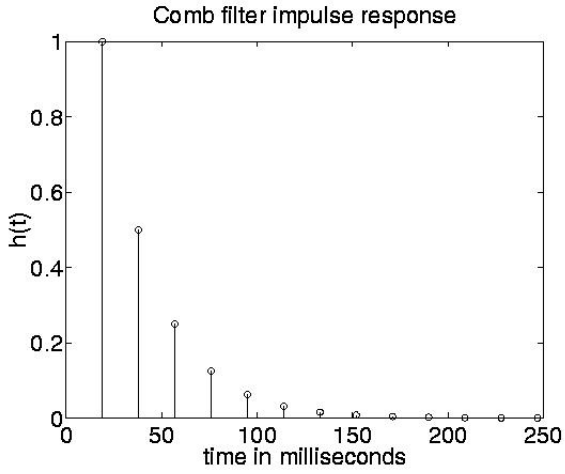


Figure 2.4: IIR Comb Filter Impulse Responses

Schroeder [30] invented the classic Schroeder reverberator that uses a combination of comb filters and all-pass filters to produce an impulse response that more nearly resembles the random nature of a physical reverberant environment.

In [30], Schroeder used four IIR Comb Filters and 2 Allpass filters. The comb filters are connected in parallel to minimize spectral anomalies (frequencies passing through one filter, while being attenuated by another), and allpass filters are connected in series because the phase distortion they produce would result in a non-uniform amplitude response were they to be connected in parallel. The purpose of the comb filters is to control the length of the reverberation, and the purpose of the allpass filters is to control its intensity. Consequently, the delay (loop) times of the comb filters will be much longer than the delay times of the allpass filters. It is important to choose delay times that are relatively prime to one another (no common

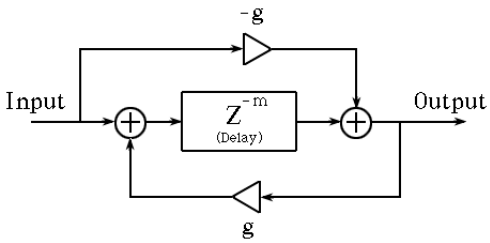


Figure 2.5: All-Pass Reverberator

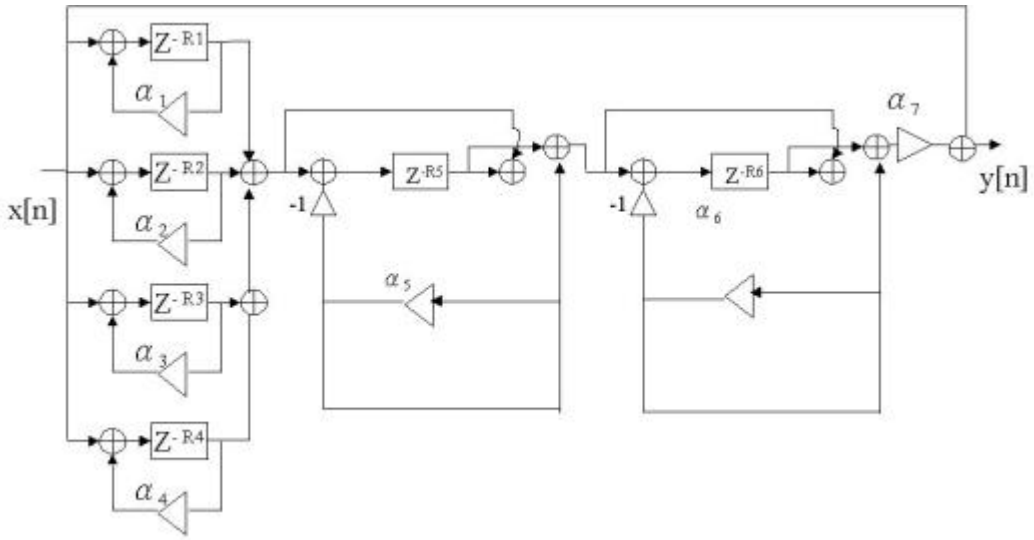


Figure 2.6: Schroeder Reverberator

divisor) because echoes will coincide with increased amplitude at multiples of common divisors. Typical values are 29.7, 37.1, 41.1 and 43.7 milliseconds [41]. As for the all-pass filters, the typical values are respectively 96.83 and 32.92 milliseconds for the delay times, and 5 and 1.7 milliseconds for the reverberation times.

There are some inevitable problems in this approach. There are peaks in the spectrum, so there is not always of flat response in the reverberation. This can be minimized by using delay times which are prime to one another. Also, the reverberation, while it looks random, is still somewhat periodic, which the ear can track. This wouldn't occur naturally in reverberant spaces, so it sounds wrong to us.

Schroeder [30] suggested that the sound of the reverberation can be improved by making the gain in the feedback loop of the IIR Comb Filter, g_1 g_4, frequency dependent. When using low pass filters for this purpose, it will attenuate high frequencies more quickly than lower ones, which closely emulates the reverberation properties of

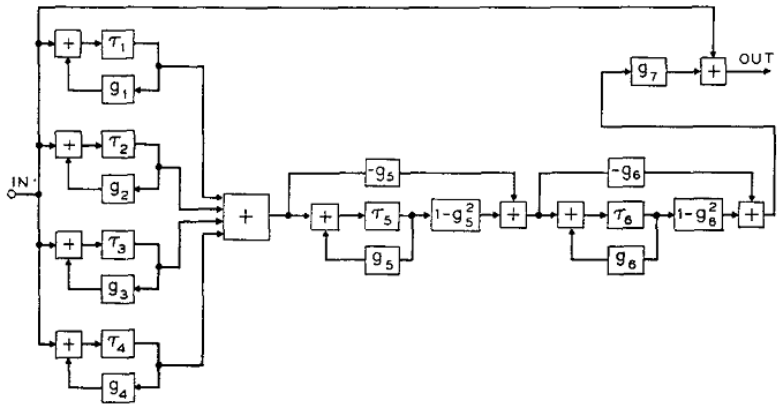


Figure 2.7: Schroeder Reverberator in The Original Paper [30]

concert halls.

The ground-breaking work by Dr Scheroder has been the cornerstone of many modern reverberators. A close variant proposed by Moorer [38] uses six(6) comb filters in parallel followed by a single all-pass filter. To simulate the attenuation of higher frequencies by the air, he incorporated a first order low-pass filter in the loop of each comb filter. By carefully selecting the delay and gain parameters, Moorer achieved a good-sounding, smooth artificial reverberation which eliminated some of the problems that the Schroeder reverberators exhibited.

Built upon a number of key improvements made to Schroeder reverberator, the Feedback Delay Network (FDN) based reverberator developed by Jot [42] [28] [43] is the most widely used one because it is the core of Creative's EAX and OpenAL implementation that shipped with millions of Sound Blaster branded PC sound cards [44].

First introduced by Gerzon [45] and then refined by other researchers, FDN contains a multichannel delay line with a feedback matrix. Such a system must be unitary, meaning that its output signals must preserves the total energy of all input signals. An early realization of FDN [46] consisting of four delay lines and a feedback matrix is shown in Fig 2.8. The feedback matrix allows the output of each delay line to be feed back to each delay input, with the matrix coefficients controlling the weights of these feedback paths. The structure can be seen as a generalized parallel comb filter structure in Schroeder Reverberator, which would arise using a diagonal feedback matrix. This structure is capable of generating much higher echo densities than the parallel comb filter, given a sufficient number of non-zero feedback coefficients and

incommensurate delay lengths [8].

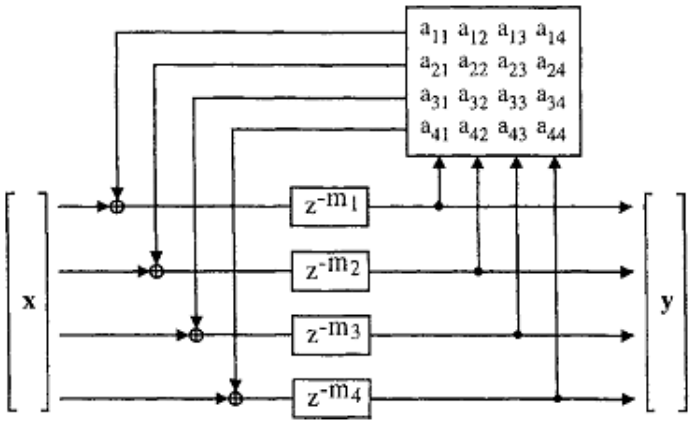


Figure 2.8: Four Channel FDN

Jot [42] [28] [43] developed a systematic FDN design methodology allowing largely independent setting of reverberation time in different frequency bands. Using Jot's methodology, FDN reverberators can be polished to a high degree of quality, and they are presently considered to be among the best choices for high-quality artificial reverberation [47].

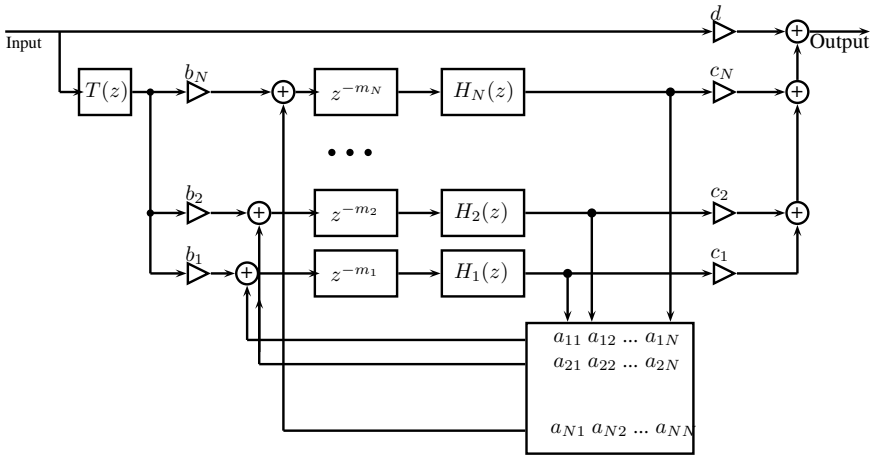


Figure 2.9: FDN reveberator by Jot

The selection of the feedback matrix plays a critical role in designing any FDN based reverberation. Experiments have been carried out and several schemes that

satisfy the "unitary" condition have been put forward, such as triangle matrix, Householder matrix [42], Hadamard Matrix[46], etc. More instructively, Jot has listed some principles of matrix selection, among which the most important ones are:

1. The matrix A should have no null coefficients, so that the recirculation through multiple delays produces a faster increase of the "echo density" along the time response. Take the classic Schroeder Reverberator as example, FDN with diagonal matrix generates reverberation of low density, for which we need to connect a few all-pass filters in series to ensure the density quality

2. To speed up the convergence towards a Gaussian amplitude distribution, the "crest factor" of the matrix A (ratio of largest coefficient over RMS average of all coefficients) should be minimum. Ideally, all coefficients should have the same magnitude. The Crest Factor of diagonal matrix is considerably big.

The feedback matrix structure of Fig. 2.9 can be improved to simulating various absorption in rooms by attaching a lowpass filter $H_n(z)$ at the end of each delay line. These low pass filters keep the reverberation time longer for low than for high frequencies.

Another interesting addition in Jot's reverberation is a low-order filter $T(z)$ applied to the non-direct signal. Called a *tonal correction* filter by Jot, this filter serves to equalize modal energy irrespective of the reverberation time in each band. In other words, if the decay time is made very short in some band, $E(z)$ will have a large gain in that band so that the total energy in the band's impulse-response is unchanged. This is another example of orthogonalization of reverberation parameters: In this case, adjustments in reverberation time, in any frequency band, do not alter total signal energy in the impulse response in that band.

By manipulating various parameters in the system, the reverberator structure proposed by Jot has two distinguishing properties [8]

1. It can be designed with arbitrary time and frequency density while simultaneously guaranteeing absence of tonal coloration in the late decay.

2. It can be specified in terms of the desired reverberation time and frequency response envelope.

Jot [48] also proposed an approach for realizing a recursive digital display network capable of simulating in real time the perceptively relevant characteristics of the reverberation decay in a room. This analysis/synthesis method demonstrated that it

possible to imitate the late reverberation of a given room by optimizing some of the reverberant filter's parameters. The analysis phase is based on a time-frequency representation of the energy decay. The energy decay relief is then used as a spectral development of the integrated energy decay curve introduced by Schroeder. By adjusting the filter coefficients of his FDN reverberator, the response of the system can be modified to approximate the time-frequency characteristics of the target room. Later, Smith and Rocchesso [49] found that FDN can be viewed as a spacial case of a digital waveguide network (to be discussed in 2.2), making its connection to the physically based approaches.

Another interesting alternative has been proposed [50] that is to produce artificial reverberation in the frequency domain, using spectral magnitude decay. This method involves accumulating the magnitudes of the short-time Fourier transform, based on the desired decay time as a function of frequency. This approach has several advantages such as it requires less memory than time-delay based methods such as FDN and provides independent control of the reverberation energy and decay time in each frequency bin. However, it requires higher computational cost.

2.1.3 Complex Scenes

Another popular topic in the area of spatial audio rendering in which the perceptual audio rendering is utilized is the reproduction of complex acoustical scenes. In many applications such as gaming and film production, a large number of sound sources need to be rendered simultaneously. Some audio acceleration hardware found in PCs and gaming consoles are capable of doing basic 3D rendering up to a few hundreds of sound sources. For example, EMU20kx audio chip inside Sound Blaster X-Fi line of sound cards can handle 128 3D voices, whilst XBOX 360 can handle 256 audio voices. When the number of sound sources increases or more than basic rendering on each source is required, the accelerating hardware become powerless. Given current computing technology, simplification of complex acoustic scenes based on perceptual criteria is the only viable option.

Inspired by the occlusion culling and level of detail algorithms widely used in computer graphics, Tsingos *et al* [32] proposed a spatial audio rendering pipeline that is capable of clustering sound sources and eliminating insignificant ones based on geometry and perceptual saliency. This method reduces the total number of effective 3D sound sources and in turn reduces the computational cost greatly. Based on

the fact that there is significant psycho-acoustic evidence that rendering each sound source individually might not be necessary due to limits in our auditory perception and localization accuracy [51], it estimates the perceptual saliency of each sound sources and then uses this information to drive their culling and clustering algorithms. Each cluster is represented by one impostor sound source, positioned using perceptual criteria. Spatial audio processing is then performed only on the impostor sound sources rather than on every original source thus greatly reducing the computational cost.

Perceptual criteria can also be used to reduce the amount of data needed to store and transmit a large amount of acoustic signals. Upon the assumption that "*the more the energy of a source in the sum signal dominates in a critical band the more perceptually relevant are the localization cues in that band. If several sources share the same localization cues they are treated as one source*".

Faller and Baumgrate [37] proposed a scheme for simultaneous placement of a number of sources in auditory space that is based on an assumption about the relevance of localization cues in different critical bands. Given the sum signal of a number of sources and a set of parameters (side-information) their scheme is capable of generating a binaural signal by spatially placing the sources contained in the monophonic signal. This scheme can be expanded to produce stereo sum signal.

Overall, perceptual based methods are very attractive because of their flexibility and efficiency.

2.2 Physical Approaches

Physical approaches model physical principles of sound propagation in air and reflections from boundaries. If the models are accurate and sufficient, a very high quality rendering system can be based on them. To achieve this goal, all of the following three components must be modelled properly:

1. Sound sources;
2. Transmission medium (room acoustics);
3. Receivers(Listeners).

All three components are discussed in the following sections, with emphasis on the room acoustics modeling and reproduction.

2.2.1 Sound Source Modeling

Sound sources are presented in recorded or synthesized digital audio data. Almost all sound sources display certain directivity pattern. For instance, most musical instruments have radiation patterns, as do human speakers. These directivity patterns are often frequency dependent. An accurate model of the sound source directivity pattern would require measuring the sound pressure from all directions across the entire audible frequency range. However, typically sound sources radiate more energy to the frontal hemisphere whereas sound radiation (especially its low frequency energy) is attenuated when the angular distance from the on-axis direction increases. Based on this observation, simple models can be devised.

DirectSound, Microsoft's gaming and multimedia API, uses sound cones to represent the source directivity [52]. Sound cones are made up of an inside (or inner) cone and an outside (or outer) cone. At any angle within the inner cone, the volume of the sound is just what it would be if there were no cone. At any angle outside the outer cone, the normal volume is attenuated by a adjustable factor. Between the inner and outer cones is a zone of transition from the inside volume to the outside volume. The volume decreases as the angle increases. Fig. 2.10 shows the concept of sound cones.

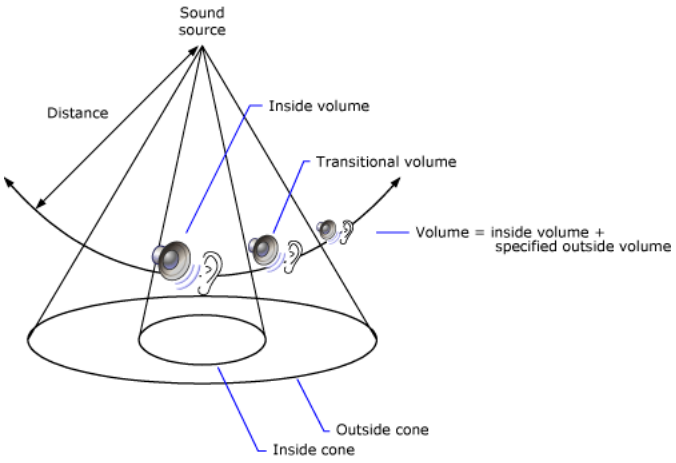


Figure 2.10: A Sound Cone in DirectSound3D

In order to model the frequency dependent directivity pattern, Savioja *et al* uses a set of angle-dependent digital filters designed from measurements of the target sound source [53], as depicted in Fig. 2.11.

Another important aspect of source modeling is the distribution pattern of the source. In most existing spatial sound rendering systems, the sound sources are con-

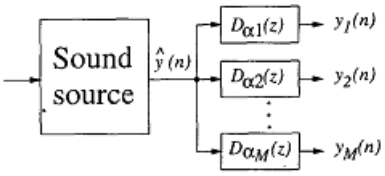


Figure 2.11: Directional Filtering [53]

sidered point-sources, which means they appear to emanate from a single direction in 3D auditory space. In real-world conditions, many sound sources generally approximate the behaviour of point sources. However, some sound-emitting objects radiate acoustic energy from a finite surface area or volume whose dimensions render the point-source approximation unacceptable for realistic 3D audio simulation. Such sound-emitting objects may be more suitably represented as line source emitters (such as a vibrating violin string), area source emitters (such as a resonating panel) or volume source emitters (for example a waterfall) [54]. A common approach to simulate these distributed source is to use multiple-point-source approximation [53].

2.2.2 Receiver(Listener) Modeling

Binaural Systems

As discussed in the previous chapter, depending on target reproduction system, the modeling techniques vary greatly. The binaural reproduction over headphone feeds audio signals directly to ears, and therefore requires modeling the characteristics of entire signal path along which the sound waves travel to human ears, include the interaural time difference (ITD), the interaural intensity difference (IID) and the direction(angle)-dependent filtering due to the pinnae, head and torso of the listener. The combined representation of these static localization cues is often referred to as the head-related transfer function (HRTF). Two main categories of approaches exist for realizing HRTF, namely filter design based approaches and structural model based approaches.

Filter design based approaches rely on measured HRTF databases such as The CIPIC HRTF Database[55] and the LISTEN HRTF Database[56]. The measured head related impulse response (HRIR) can be used to render binaural audio when convolved with the input sounds. However, this would require storage for very large amount of HRIR data to represent all possible sound arriving directions with sufficient

resolution. This also imposes intensive load on the computing hardware. For example, Dolby Headphone™uses 7,000-tap FIRs to simulate the HRIRs [57]. Therefore, it is desirable to seek very efficient filter design and implementation methods, as well as approximation and smoothing techniques that can reduce the number of HRIR measurements needed. Significant reduction can be achieved by taking into account the properties of the human auditory system, that is, the human auditory system processes information on a logarithmic amplitude and nonlinear frequency scale. One example of such simplification method was proposed by Jot [58] where the HRTFs are preprocessed using auditory smoothing and an IIR approximation is designed in the warped frequency domain using standard IIR design approaches. A thorough review of HRTF filter design can be found in [59] and more recently [12]. One main setback of the filter based HRTF modeling is that HRTF varies significantly from person to person. To achieve convincing elevation effects, the HRTF (or HRIR) must be measured for each listener, which is both time consuming and extremely inconvenient.

Rather than measuring HRTFs and simulating them using digital filters, another approach is to design **structural models** to capture the important physical features that affect sound propagation from outside to human ears. Brown and Duda [60] proposed a simple signal processing model of the HRIR for synthesizing binaural sound from a monaural source. The components of the model have a one-to-one correspondence with the shoulders, head, and pinnae, with each component accounting for a different temporal feature of the HRIR. Specifically, this model contains separate components for azimuth (head shadow and ITD) and elevation (pinna and shoulder echoes). Head shadow filters are implemented by IIR and echo filters by FIR, as in Fig. 2.12. The main advantage is that the simplicity of the filters enables inexpensive real-time implementation without the need for special DSP hardware and significantly reduces the storage size. Furthermore, the parameters of the model can be adjusted to match the individual listener in order to produce individualized HRTFs. A3D audio rendering system developed by Aureal is believed to base their HRTF synthesis on the structural model [29].

In addition to modeling and incorporating HRTF, binaural reproduction over loudspeakers also requires modeling and compensating the acoustical path from the loudspeakers to the ears. This is known as crosstalk cancellation [5]. It is possible to build an elaborate digital filter, called a "crosstalk canceller", to eliminate crosstalk. As depicted in Fig. 2.13, if $Hi(z)$ and $H_c(Z)$ represent the ipsilateral and contralat-

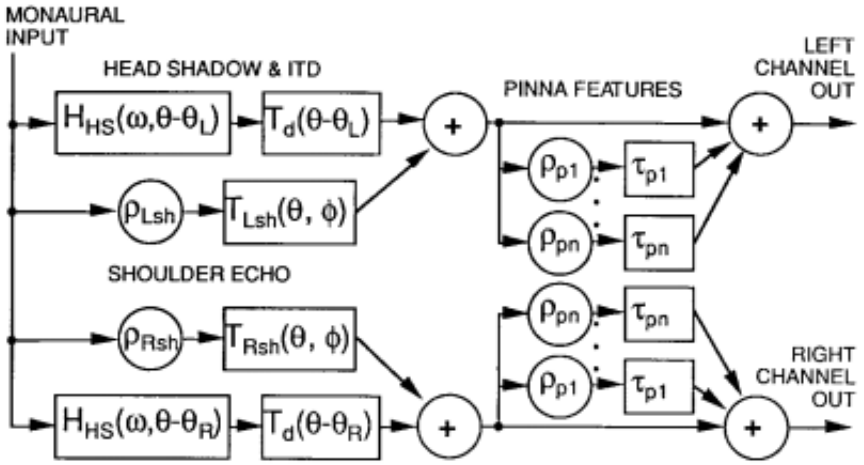


Figure 2.12: A Structural Model of HRIR [60]

eral loudspeaker-to-ear transfer function, the purpose of the crosstalk cancellation filters are to invert the effect of $H_i(z)$ and $H_c(Z)$. The difficulty in designing good sounding inverse filter and the well-known problem of the "sweet spot" in loudspeaker listening of binaurally processed audio reduces the usability of crosstalk cancellation techniques [53]. For the same reasons, binaural reproduction over loudspeakers has not developed into mature technology that can be deployed widely. An excellent review and discussion on this topic can be found in [5].

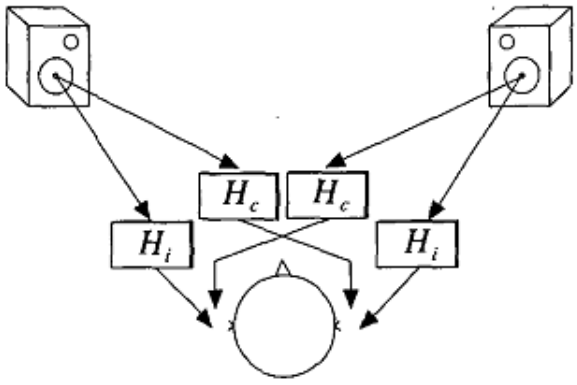


Figure 2.13: Crosstalk

Multichannel Speaker Systems

A natural choice for creating a virtual acoustical space is to use multiple loudspeakers in the reproduction. With this concept, the problem of generating spatial auditory cues is reduced to the placement of the loudspeakers and the manipulation of the signals that feed the loudspeakers. One obvious benefit of multichannel speaker systems is that they are capable of creating larger listening areas and producing a more stable localization effect. Another advantages is that in multichannel loudspeaker reproduction, no binaural modeling is required, and thus it is computationally less demanding.

The vector base amplitude panning (VBAP) mentioned previously is one way of manipulating of the signals that feed the loudspeakers. It is not physically accurate and can not be used for the signals where the delay and timing information is required, such as reflections. It is, however, possible to extend the VBAP to include delay as a pre-processing step.

Wave field synthesis (WFS) [61] is by far the most accurate modeling and reproduction method that is capable of reconstructing an 'acoustically correct sound field'. WFS is based on Huygens' Principle, which states that any wave front can be regarded as a superposition of elementary spherical waves. Therefore, any wave front can be synthesized from such elementary waves. In practice, a computer controls a large array of individual loudspeakers and actuates each one at exactly the time when the desired virtual wave front would pass through it. Contrary to conventional spatialization techniques such as stereo, the localization of virtual sources in WFS does not depend on or change with the listener's position. However, there are a number of basic and practical problems which are summarized below.

1. High cost versus aliasing. A large number of individual transducers must be very close together. Otherwise spatial aliasing effects becomes audible. This is a result of having a finite number of transducers (and hence elementary waves). In order to deceive human ears that no spatial aliasing is occurring, the loudspeakers must be placed no more than 10˜15 centimeters apart.

2. Spatial interference from the listening environments. Since WFS attempts to simulate the acoustic characteristics of the recording space, the acoustics of the rendition area must be suppressed.

3. Other problems include truncation effects [62] and undetermined adversary ef-

fects on perception [61].

WFS is a relatively new technology that showed some potential. However, significant amount of effort is needed to develop mature technologies around WFS and achieve large acceptance [62].

Dynamic characteristics such as head movements and interaction between the reproduced sounds and listening environment are also important factors we need to consider. Techniques such as head tracking [5] and room equalization [63] have been proposed to deal with these problems. Further discussion is beyond the topic of this thesis.

2.2.3 Room Acoustic Modeling: Geometric Methods

In this thesis, geometric based methods refer to the kind of methods that construct a computational representation of the acoustic space being modeled and derive the sound propagation paths from this representation. Fig .2.14 shows the simplest 2-D example that contains a rectangular room (with 4 walls), one sound source and one listener. A more complex scenario may include acoustically significant objects (that can cause occlusion and diffraction), and it may also include acoustical characteristics of the surfaces such as reflectance, roughness/unevenness and edge conditions.

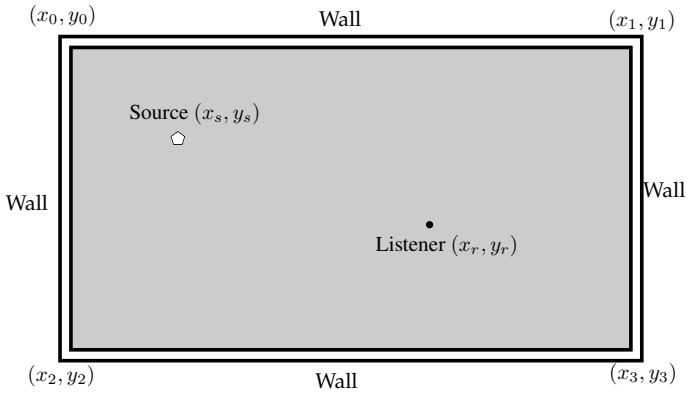


Figure 2.14: Simple 2-D Geometric Model

Based on ray theory, the geometric based methods are analogous to the tracing of light rays, a method of following the paths light takes from source to receiver. Rays are constructed from the sound source modeled as obeying the wave equation and are

tracked through the virtual environment. Mathematical models are used to approximate the filters corresponding to source emission patterns, atmospheric scattering, surface reflectance, edge diffraction, and receiver sensitivity for sound waves traveling along each path. The more challenging step of geometry based methods is to find all the propagation paths. Several solutions to this problem have been proposed such as the ray-tracing method, the image source method and the beam tracing method.

Ray tracing methods find reverberation paths between a source and receiver by generating rays emanating from the source position and following them through the environment until an appropriate set of rays has been found that reach a representation of the receiver position. The main advantages are that it models all types of surfaces as well as scattering and it is simple to implement [64]. However, it is subject to aliasing and depends on the receiver position which means once the listener moves, all the propagation paths must be regenerated.

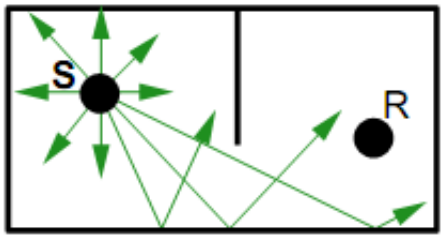


Figure 2.15: Ray Tracing

A3D audio API [29] developed by Aureal in conjunction with clients such as NASA, Matsushita and Disney, uses a proprietary Wavetracing algorithms [1] to parse the geometry description of a 3D space and to trace sound waves in real-time as they are reflected and occluded by passive acoustic objects in the 3D environment. The geometry description also includes wall surface materials for simulating proper wall absorption. In Aureal's implementation of A3D using the Vortex A3D Silicon Engines, reflections are rendered as individually imaged early reflections and late reverberation as field reflections.

Image source modeling of wave propagation provides a simpler, yet limited solution. The concept of image sources is very straightforward: for each reflective surface, a virtual sound source is constructed by mirroring the sound source across the surface.

[1] No further information is available, but it is widely believed this is based on raytracing algorithms.

Therefore, as the listener travels through the space, the reflection is modeled by the sound emanating from the virtual source (taking into account attenuation based on boundary characteristics). The primary advantage of image source methods is their robustness in that they guarantee that all specular paths up to a given order or reverberation time are found. Image source methods are very efficient in the special case of box-shaped environment due to the rectilinear symmetries of a box [64]. However, this method models only specular reflections and becomes inefficient in oddly shaped spaces [65].

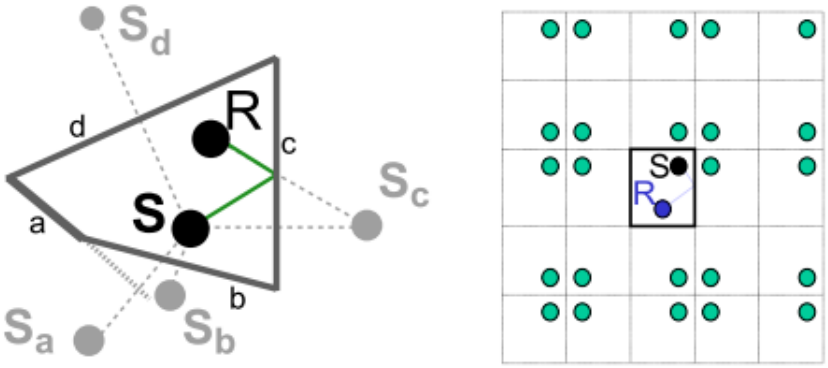


Figure 2.16: Image Source Method (a) irregular room (b) efficient expansion of box-shaped room

Beam tracing is closely related to, and more efficient than ray tracing. Instead of calculating the paths of a large number of sound rays, these rays are collected into three-dimensional beams; 3D space is subdivided into convex polygons based on the propagation and reflection of sound from a source [66]. One of the main advantage of beam tracing is that beams can be precomputed, during an off-line phase, and stored in a data structure (e.g., a beam tree) for later evaluation of reverberation paths at interactive rates. Its disadvantages include that it is not able to handle curved surfaces and it requires polygon sorting and intersection which may become the efficiency bottleneck [64].

Once geometric propagation paths have been computed, they are combined to form filters for spatializing a sound signal. The challenge now is to model the attenuation and scattering of sound as it travels along each path, taking into account distance delay and attenuation, atmospheric scattering, reflectance functions and diffraction models. These components and their realization are discussed in the latter chapters.

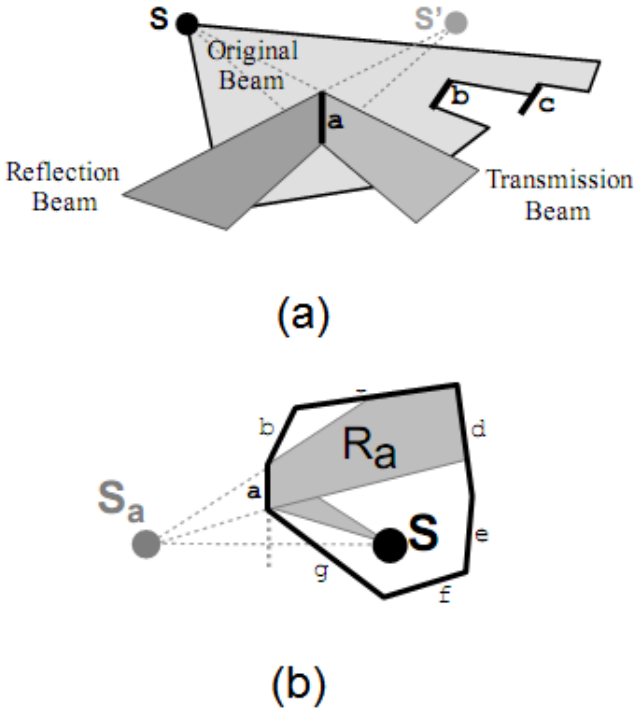


Figure 2.17: Beam Tracing Method (a) principle (b) culling invisible virtual sources

In theory geometric base methods can be highly accurate, as they model sound based on wave propagation as is known in physics. However, they make the assumption that sound wavelengths are significantly smaller than the size of obstacles, and thus they are valid only for high-frequency sounds. In addition, when dealing with higher order of reflections, even the most efficient geometric method involves an extraordinary number of computations and is not plausible for most interactive systems. Tsingos [67] proposed a solution to migrate the heavy geometrical computation from run-time to off-line by running ray tracing algorithm for a few key locations offline and storing parametric reverberation effects for interactive rendering. This inevitably sacrifices the flexibility and complicates system design. Another disadvantage of geometry based approaches is that it often fails to reproduce high fidelity late reverberation because the errors from inaccurate and/or incomplete path finding and inaccurate reflection modeling accumulate exponentially[68]. For example, if the wall reflectance is off by 10%, after being reflected 10 times, the error will be $110\%^{10} = 259\%$!

2.2.4 Room Acoustic Modeling: Room Impulse Response Based Method

The room impulse response (RIR) based method, also known as convolution reverberation, refers to digitally simulating the reverberation of a physical or virtual space by convolving a pre-recorded (and normally dry) audio sample with the room impulse responses (RIRs) of the space being modeled.

The two most attractive merits of RIR based method are their simplicity and very high fidelity. Reverberation is often seen as a time invariant and linear effect [8]. It is time-invariant because in a given physical and virtual space, the reverberation characteristics do not change over time. It is linear in that it obeys scaling and superposition principles. A time-invariant linear system (LTI system) can be completely characterized by its impulse response. This is why the ultimate objective criterion of judging the quality of perceptual based or geometric based reverberation is to compare the generated IR with the actual (measured) IR to see if they match [69]. The problem of creating the reverberation effect is equivalent to filtering the input signal with the proper impulse response. This becomes a straightforward filtering problem that is able to produces high fidelity reverberation because the impulse response fully represents the acoustical characteristics of the target configuration [2]. Mathematically, this process can be represented as

$$y(t) = \int_0^\infty h(\tau)x(t-\tau)d\tau \tag{2.1}$$

where $y(t)$ is the signal at receiver (microphone or listener), $x(t)$ is the signal emitted by the sound source and $h(t)$ is the impulse response. The discrete time version is

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k) \tag{2.2}$$

where $h(k)$ of size N is the truncated version of the actual impulse response. N must be selected properly based on the target space to avoid losing information. For example, a proper IR of a concert hall may exceed 3 seconds, which means $N = 132,300$ if 44.1KHz sampling rate is used.

The main obstacles that have prevented the convolution reverberation from reaching mainstream adoption have been the difficulties in obtaining the RIRs and the for-

[2]The configuration includes the listener position, the source position and the space.

bidding computation cost. The direct measurement of the RIRs uses an impulse-like signal source and records the response from a receiver position. However, impulse like signals such as a gunshot or a hand clap do not contain sustained energy in order to generate sufficient signal noise ratio (SNR) for late reverberation [70]. Other measuring methods have been proposed that use other types of excitation signal such as maximum-length sequence (MLS) and frequency sweep. There methods dramatically improve the quality of the acquired impulse responses. The details of these methods are discussed in the next Chapter.

Another main obstacle has been the forbidding computation cost. Given a $N = 132,300$ IR, the direct convolution requires 132,300 multiplications and additions for each input audio sample. With advances in computing hardware and the use of FFT-based fast convolution, this obstacle no longer exists. For example, a properly optimized fast convolution reverberation of 3-second IR with stereo (2 channel), 44.1KHz sampled audio input takes about 6% CPU of one of the eight SPU cores in the CELL processor [3].

In the past decade, the RIR based method is rapidly gaining popularity as a powerful sonic tool [72]. There are many software tools dedicated to this purpose, such as Altiverb by AudioEase, Christian Knufinke SIR2 and Tascam GigaPulse (see [73] for more listing). There are also many IR libraries available to use with the tools such as Samplicity[74] and Open Impulse Response Library[75].

However, there are two major problems remaining. The first one is that one IR correspond to one listener-receiver-space configuration. This means when a listener or a receiver moves to a different position, it is a new configuration and needs a new IR. Therefore, the target space must be spatially sampled sufficiently and a separate IR must be measured and stored for each sample point. Many convolution reverberation software and libraries contains a number of IRs to represent a predefined set of source positions, for example Fig. 2.18 shows the sample points used in an AltiVerb example. For these approaches, it is desirable to *sample* the acoustic space adequately by measuring the impulse responses at a large number of positions. This is because each impulse response is unique and contains new information, especially new directional information. In the later chapters, we will show that by combining with geometry based approaches, only one representative impulse response is required for an acoustic space or a section within an acoustic space that has a distinct late

[3]The number is based on the benchmark obtained during my work-term at Electronic Arts Canada in Burnaby BC. For more information on the CELL processor, please refer to [71].

reverb characteristics.

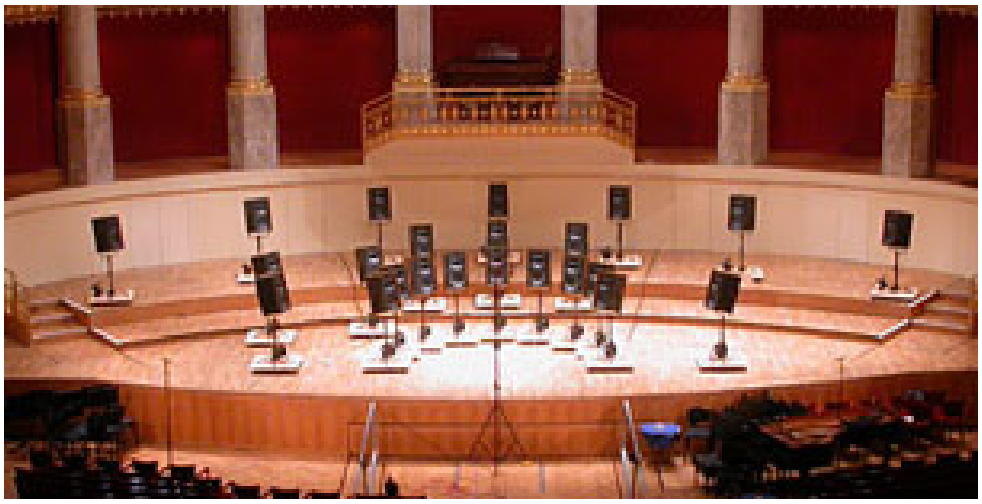


Figure 2.18: Sampled Orchestra Stage Positioning [76]

The other problem is that there is no easy way of interpolating IRs. Consider we have IRs for source position A and position B, $h_a(n)$ and $h_b(n)$, a smooth transition from $h_a(n)$ to $h_b(n)$ is required for interactive applications. However, as illustrated in Fig. 2.19, an interpolation in the normal sense is not able to produce the smooth transition, which is required for rendering moving sources.

Vienna MIR [77] claims to allow for "seamless interpolation" of each sample point within the available areas of the room ("Hot Spots"). However, it uses a specially formatted IR library that is not inter-operable with other software and libraries, also the interpolation is limited to very small areas around the hot spots.

Another limitation of IR based method is that, while it is well suited for reproducing the acoustic impression of the target spaces, it is not as flexible as perceptual methods or geometric based methods for rendering virtual spaces. Perceptual methods or geometric based methods both have a set of adjustable parameters that can be used to create or tweak a virtual space, such as various filter coefficients, room dimensions and wall material reflectance. IR based methods rely on a set of predetermined RIRs, and it is not straightforward to create a virtual space by modifying the RIRs. However, limited control of produced acoustic characteristics by modifying RIRs is possible. For example, Altiverb [76] offers control over reverberation time by applying an exponential gain curve to the impulse response, and room size by transposing room modes and resonances, tightening/spreading early reflections and

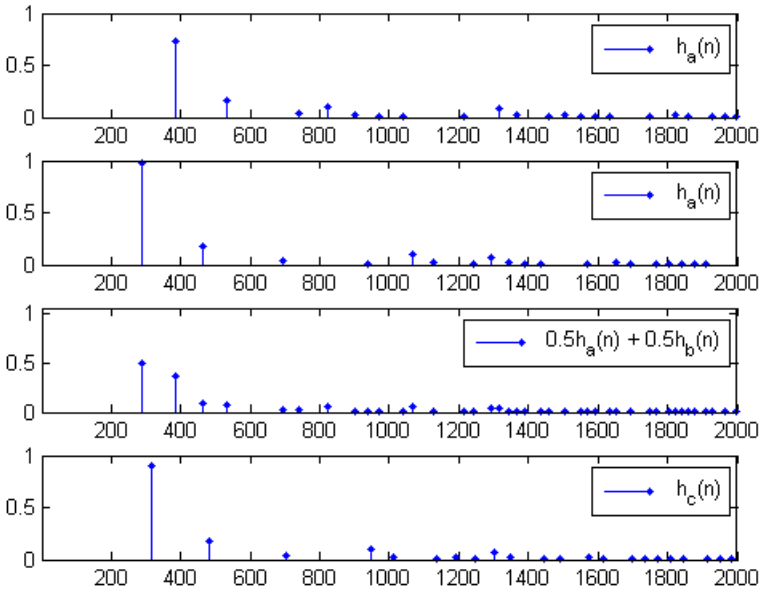


Figure 2.19: Interpolation between IRs. $h_c(n)$ is the IR of the middle point between A and B. $0.5h_a(n) + 0.5h_b(n)$ is clearly not a correct approximation.

shortening/lengthening the reverberation tail.

In summary, IR based methods are straightforward and yield very high quality reverberation. For these reasons, they have been very popular in studio production. However, they are not able to handle source or listener movements and have very limited use in interactive applications such as gaming.

2.3 Hybrid Methods

From the above discussion, it is clear that no single solution is capable of producing high quality spatial rendering while still being flexible in order to handle changes in the acoustical scene as well as being efficient enough to be used in a wide variety of applications. Several hybrid solutions have been proposed that try to combine the benefits of different approaches and to overcome the individual disadvantages.

The DIVA project [31] [2] developed at Helsinki University of Technology uses a modified image source method to find early reflections and uses a set of low order IIR filters to simulate air and wall absorption. It also incorporate an late reverberation unit that consists of several parallel feedback loops which contain a delay line, a comb-

Figure 2.20: Vienna MIR

allpass filter, and a lowpass filter. The DIVA system, as depicted in Fig. 2.21, is an efficient system that is able to generate direct sound and early reflections accurately, which is crucial for sound source localization. However, the late reverberation is based on a perceptual model and its fidelity is often questionable.

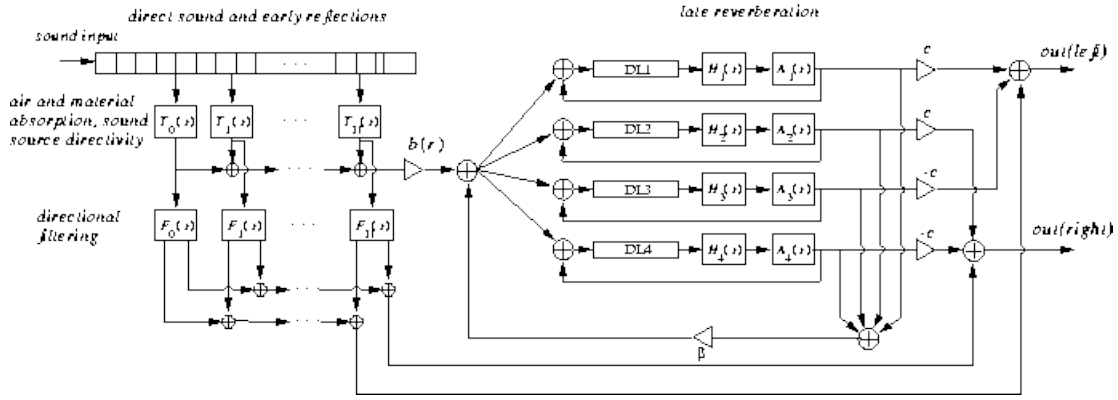


Figure 2.21: DIVA System[31]

A similar system was proposed by Rindel *et al* in [78]. In this system, early reflections are represented by time delay and energy level. It is found that in most well-behaved rooms a number less than 40 seems to be sufficient for obtaining a realistic audible impression of the simulated room. Late reflections are described

statistically by calculated reverberation time as a function of frequency. The settings of an electronic reverberation processor are adjusted according to the calculated values in a number of frequency bands.

Several wave field synthesis methods [79] [80] combine model based method for rendering moving sources and RIR based method for reverberation, as shown in Fig. 2.22. However, there is no evidence that two approaches are integrated in a consistent manner that ensures the clarity of the overall acoustic impression.

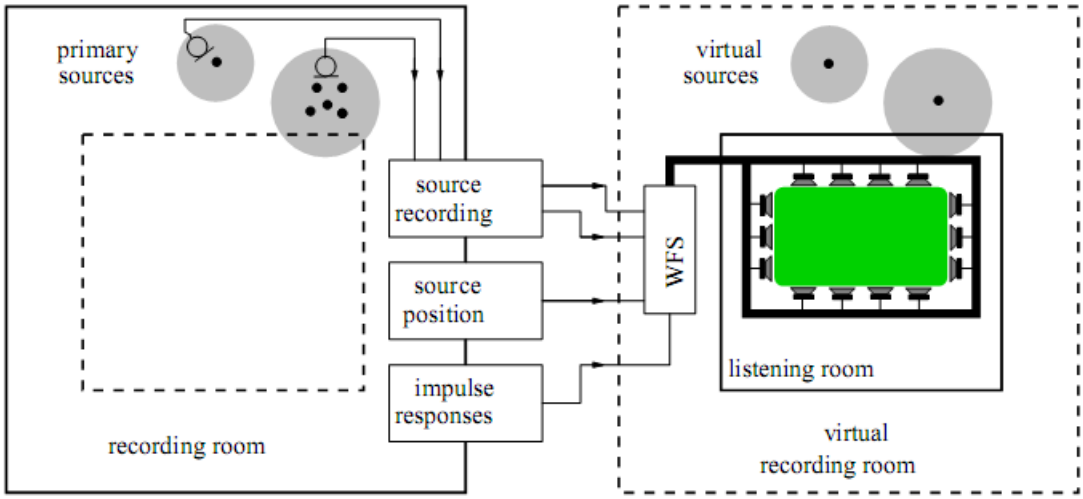


Figure 2.22: A Typical Wave Field Synthesis System[79]

Rayverb [81] developed by Prosoniq is an interesting alternative. It uses a so-called "inverse raytracing" to analyze a pre-recorded impulse response, then try to identify the size and composition of the acoustic space where the RIR was recorded in and the position of a sound source within that space. Once these calculations are made, the room model created can be used to predict what the sound source might sound like if it was repositioned within the same reverberant space.

The spatial rendering solution discussed in this thesis is also a hybrid method that is built upon multichannel measured RIRs (MMRIRs). The MMRIRs is analyzed to determine a set of parameters that affect the acoustic characteristics of the target room. A image source room model that incorporates such parameters is used to find and render the direct sound and early reflections. Segments of the MMRIRs are selected and modified in such a way that convolutive reverberation can be used to render late reverberation. Comparing to the DIVA system, the proposed solu-

tion offers a potentially much higher reverberation quality by using the measured RIRs, at the same time retains the same flexibility of freely placing and moving the sources and the listener. The current design targets multichannel speaker systems for reproduction. However, it can be easily extended for binaural systems.

Chapter 3

IR Measurement and Analysis

The acoustics of a reverberant space add feeling and life to music. Many concert halls are famous for their sound quality and many recording artists go to great lengths and cost to record live performances at these venues, in order that the listener can experience the concert hall surroundings in their own living room. Applying the acoustic response of a concert hall to music recorded in a studio would save the industry a lot of money and also allow the same piece of music to be experienced at different venues [82].

As discussed in the previous chapters, under the assumption of source and receiver immobility, the acoustical space in which they are placed can be considered a linear time-invariant system characterized by an impulse response $h(t)$ [70]. The ultimate solution to this problem, from a digital signal processing perspective, is to convolve the dry musical signal recorded in a studio with the room impulse response of the target hall. There are several problems and difficulties with this approach. For instance, convolution is a very expensive computation and a measured impulse response corresponds to a single source-listener configuration. On the other hand, although the "artificial reverberators" can possibly run in real-time and are able to simulate arbitrary source-listening configurations, they often fail to create a faithful reproduction of the acoustic space. Our solution aims at bridging the gap between these two extremes, by retaining high spatial fidelity while still being flexible enough to simulate arbitrary source-listening configurations.

3.1 Measurement of the MMRIR

The first step towards this goal is to acquire sufficiently accurate RIR measurements. Measuring a room impulse response (RIR) involves recording an excitation signal to obtain the response. The response - according to the measurement method - is either the impulse response itself or an intermediate response that needs to be post processed to obtain the impulse response. Various techniques for measuring RIR have been studied [83] [84] [85] [86]. The most commonly used ones are evaluated in this thesis.

3.1.1 Signal Selection

RIR can be measured directly using an impulse excitation such as a gunshot or a handclap. However, for a big acoustic space, the impulse function does not contain sustained energy to produce high SNR in measured RIR. Moreover, signals that approximate an impulse such as an electric spark or a starter pistol do not have an ideally flat frequency response and they are not very well reproducible (the excitation signal itself is different every time thus averaging is problematic). They are therefore not used in modern measurements any more.

This lead us to indirect measurement. Given an excitation signal $x(t)$, impulse response of the device under test (DUT) $h(t)$ and the received signal $y(t)$, the cross-correlation between $y(t)$ and $x(t)$ is

$$\begin{aligned} r_{yx}(t) &= \int_{-\infty}^{\infty} y(\tau)x(\tau - t)d\tau \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x(\tau - \alpha)h(\alpha)d\alpha\right] x(\tau - t)d\tau \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x(\tau - \alpha)x(\tau - t)d\tau\right] h(\alpha)d\alpha \\ &= \int_{-\infty}^{\infty} h(\alpha)r_x(t - \alpha)d\alpha \end{aligned} \tag{3.1}$$

If $r_x(t) = \delta(t)$, i.e. an impulse function, then $r_{yx}(t) = h(t)$. This means that for any $x(t)$ that has a impulse like auto-correlation, the cross-correlation of the received signal and $x(t)$ is a good approximation of the DUT's impulse response $h(t)$.

There are several types of signal that have such a property. In order to be used in acoustic measurement, the signal must also fulfill several other requirements. The

first one is that the signal must contain all the frequencies of interest (20 Hz to 20 KHz for high quality audio). There is also an ambiguity in the requirements of the pulse duration. A short pulse will give very accurate information but will fade quickly in a large environment, conversely a long pulse containing more energy will not give such good definition but will be easily detected even after reflections off several objects.

Among the many possibilities, the two most popular excitation signals for RIR measurement are: a Maximum Length Sequence (MLS) and a chirp signal. Both options meet the frequency content requirements.

Maximum Length Sequence (MLS)

Maximum Length Sequences are periodic binary pseudo random signals that are defined as [87]

$$x[n] = (-1)^{a[n]}X \tag{3.2}$$

where $a[n] = 0$ or $a[n] = 1$ is a binary sequence. It is clear that $x[n]$ is either X or $-X$. The auto-correlation function of these sequences is

$$\begin{aligned} r_x[k] &= \frac{1}{N-1}\sum_{n=0}^{N-2} x[n]x[k+n] \\ &= \frac{X^2}{N-1}\sum_{n=0}^{N-2}(-1)^{a[n]+a[k+n]} \\ &= \frac{X^2}{N-1}\sum_{n=0}^{N-2}(-1)^{a[n]\hat{}a[k+n]} \end{aligned} \tag{3.3}$$

where $N-1$ is the period of the binary sequence $a[n]$. In MLS, $a[n]$ is generated in such as way that

$$\begin{aligned} a[i]\hat{}\,a[i+n] &= a[i+m]; \quad n \neq 0, N-1, 2(N-1), ... \\ a[i]\hat{}\,a[i+n] &= 0; \qquad n = 0, N-1, 2(N-1), ... \end{aligned} \tag{3.4}$$

This results in

$$r_x[k] = \begin{cases} 1 & \text{if } n = 0, N-1, 2(N-1), ...; \\ \frac{-1}{N-1} & \text{else.} \end{cases} \tag{3.5}$$

As shown in Fig. 3.1, except for a little DC bias, $x[n]$ has the same auto-correlation as white noise if N is large enough.

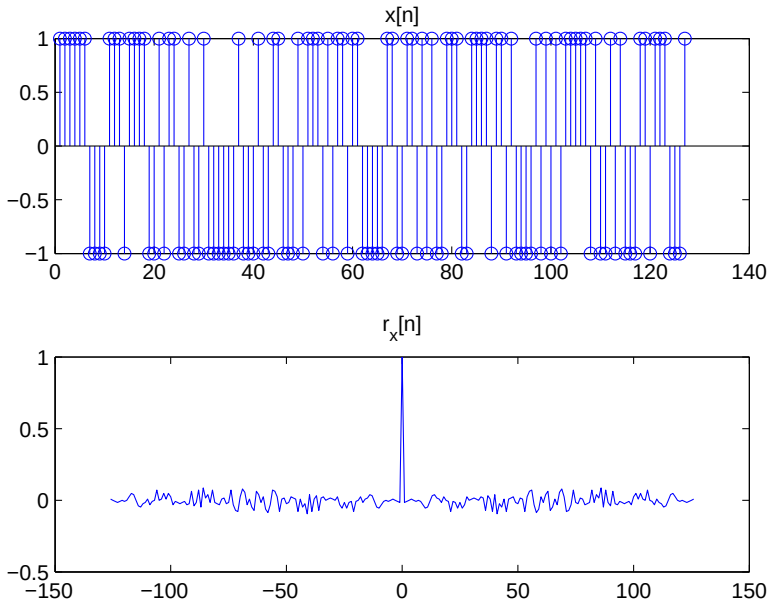


Figure 3.1: MLS and its Autocorrelation

$$r_{yx}[k] = \frac{1}{N-1}\sum_{n=0}^{N-2} y[n]x[k-n] \tag{3.6}$$

$y[n]$ is the response of the system under test to the excitation signal $x[n]$

$$y[n] = h[n] \star x[n] \tag{3.7}$$

where $h[n]$ is the impulse response of the system, and in our case, is the room impulse response (RIR). Therefore

$$r_{yx}[k] = \frac{NX^2}{N-1}\sum_{m} h[n+m(N-1)] - \frac{X^2}{N-1}\sum_{j} h[j] \tag{3.8}$$

If the impulse response is much shorter than the MLS length N-1, then the only term left in the summation we have to count is the m = 0 term, resulting in

$$r_{yx}[k] = \frac{NX^2}{N-1}h[n] - \frac{X^2}{N-1}H[0]. \tag{3.9}$$

This means that, the room impulse response is evaluated with the cross-correlation between the MLS and the signal at the reception point.

MLS works well for the noisy real-world environments and has been very popular in acoustics testing and measurements. This is because it has the lowest possible crest value [1] of 1, which makes it as noise-immune (for broadband noise) as possible [87]. However, MLS based measurement systems have the following problems [70],

1. MLS can not be used with high loudspeaker output level because of the harmonic distortion caused by high output level appears as spurious peaks within the impulse response and is very difficult to separate.

2. It is not robust to minor time-variance such as clock jitter and wind. For this reason, MLS is not suitable for large rooms and requires synchronization between receiver and transmitter.

Frequency Sweep (Chirp) Signals

To overcome the drawbacks of the MLS based measurement, frequency sweep signals can be used as excitation signals. There are two types of sweep signals, namely, linear sweep and exponential sweep.

Linear sweeps have a constant sweeping speed, upwards or downwards, from the start frequency ω_1 towards the end frequency ω_2.

$$x(t) = A \cdot \sin\left(\frac{1}{2}\frac{\omega_2 - \omega_1}{T}t^2 + \omega_1 t\right) \tag{3.10}$$

where A is the peak amplitude of the signal and T is its length. Linear sweeps have flat spectrum between ω_1 and ω_2.

Exponential sweeps have an accelerating sweeping speed and have pink spectrum.

[1] crest factor = (peak value) / (rms value)

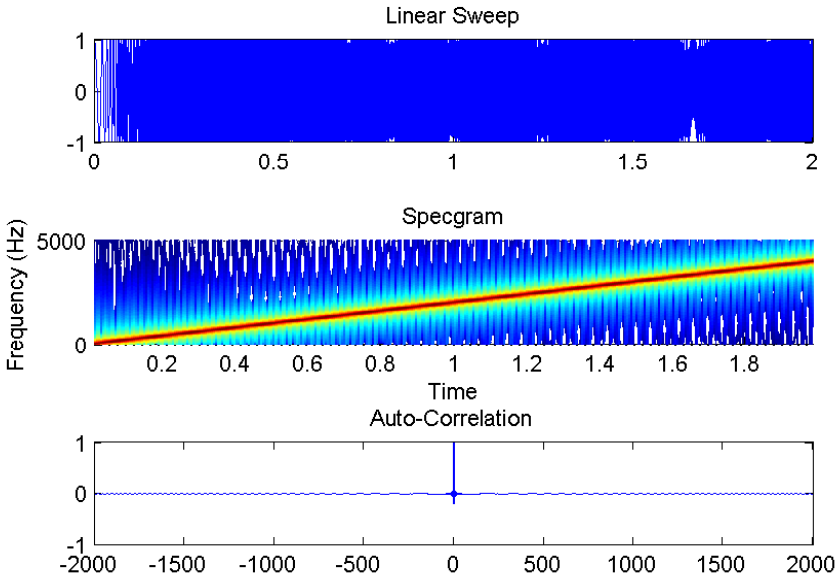


Figure 3.2: Linear Sweep

They can be synthesized in time-domain with the following formula

$$x(t) = A \cdot \sin\left(\frac{\omega_1 T}{\ln\left(\frac{\omega_2}{\omega_1}\right)}\left[\exp\left(\frac{1}{T}\ln\left(\frac{\omega_2}{\omega_1}\right)\cdot t\right) - 1\right]\right) \tag{3.11}$$

From Fig.3.2 and Fig.3.3, it is clear that both linear and exponential sweep have auto-correlation functions that are close the impulse functions. Therefore the cross-correlation between the recorded signals and these sweep signals are good approximation of the RIR.

Sine sweeps are superior in many ways compared to MLS signals. Being a continuous signal, sweep signals do not suffer from the clock jitter and other minor time variance. The harmonic distortion in the system does not degrade the quality of measured RIR as this type of distortion can be separated in time from the baseband impulse response. In the case of exponential sweeps, the time between the harmonics are constant at all frequencies [88]. Additionally, sweep signals are linear signals so they are less likely to damage the power amplification equipment [82]. Another advantage of the exponential sweep over the linear sweep is that exponential sweeps provides higher SNR in the lower frequency range because it has more energy in that

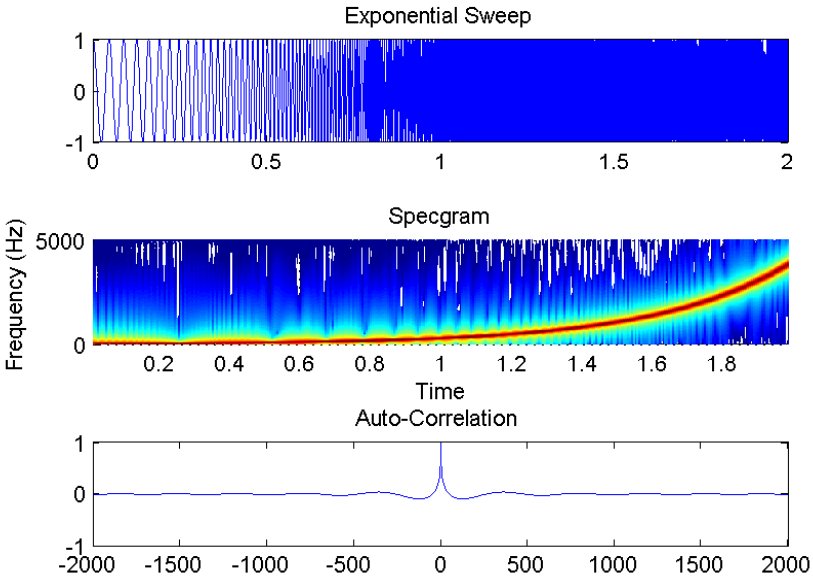


Figure 3.3: Exponential Sweep

region.

In our experiment, we used both the linear sweep and the logarithmic sweep techniques.

3.1.2 Microphone Setup

At the receiver end, we need a microphone setup that can capture the acoustic properties of the target space as accurate and complete as possible. This provides us with a set of RIRs that is representative of the space. This setup can be used to record acoustic events such as live performances so that we have a good reference when evaluating the proposed spatial rendering solution. It is also preferable that measurement from this setup can be applied to the mainstream multichannel speaker systems without additional mapping and processing. Several microphone configurations have been proposed for capturing 3D acoustic spaces, in-depth discussions can be found in [89] and [90].

Microphone Direction Pattern

It is necessary to review the typical microphone directivity patterns or polar patterns. The directivity pattern of a microphone indicates how sensitive the microphone is to different directions and the direction-dependent sensitivity can be expressed with

$$m(\theta) = a + b\cos(\theta) \tag{3.12}$$

where a and b are pattern-dependent constants and θ is the angle of the direction. Typically $b = 1 - a$. An omnidirectional microphone, obtained with $a = 1$ and $b = 0$, captures sound from all directions with equal sensitivity. A dipole microphone, also called as a figure-of-eight or bidirectional microphone, expressed by $a = 0$ and $b = 1$, captures sound from front and back of the microphone with opposite phases, and it does not capture sound from the sides of the microphone. A cardioid microphone, obtained with $a = 0.5$ and $b = 0.5$, captures sound mainly from the frontal semi-sphere, and no sound from the back of the microphone. A hypercardioid microphone, expressed by $a = 0.25$ and $b = 0.75$, captures sound from a beam in the frontal semi-sphere, and with a lower sensitivity from a very narrow beam from the back of the microphone.

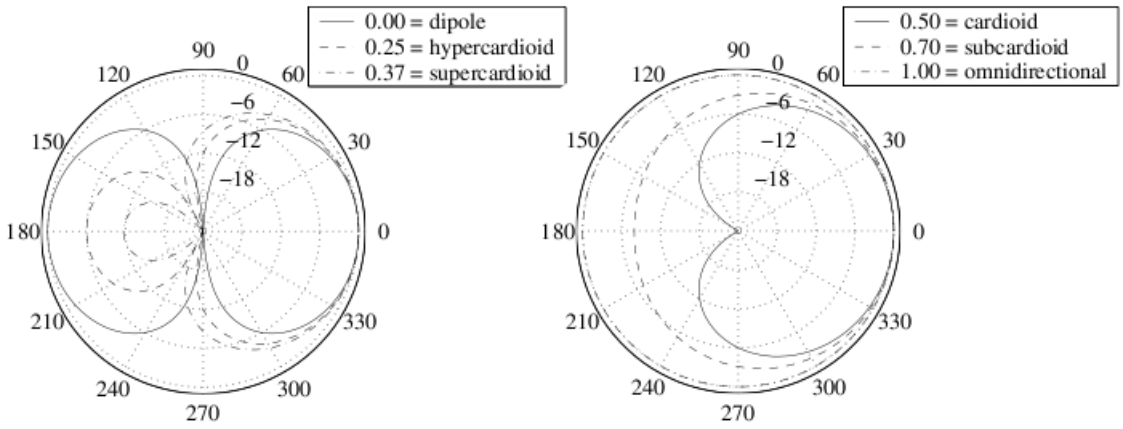


Figure 3.4: Microphone Direction Pattern

Stereo Microphone Techniques

It is important to understand the traditional stereo microphone placement techniques because the microphone array used in this thesis can be viewed as an extension of

the stereo miking technique. It is also important to understand that each technique has its unique advantages and disadvantages and no single technique works best for all scenarios.

- **Spaced Pairs.** A spaced microphone pair contains two microphones that are placed apart from each other to point to the direction of the sound source. This technique is usually used in the recording of a large instrument or a musical ensemble. With a spaced pair, the directional information of the sound is encoded in differences in both time of arrival and level of the sound signal. AB pair is a commonly used spaced pair setup that uses two omnidirectional microphones with $\theta_1 = \theta_2 = 0^o$.

- **Coincident Pairs.** A coincident microphone pair has two directional microphones that are placed in an angle with respect to each other so that their capsules are almost located at the same point. In practice, the microphones are usually placed on top of or next to each other. With this positioning, the direction of arrival is encoded only by the level difference and the issue of phase difference between the microphones is not present. Some widely used coincident pairs are XY, Blumlein and MS pairs.

- **Near Coincident Pairs.** A near coincident microphone pair can be viewed as a compromise between a spaced pair and a coincident pair as the microphones are positioned close enough to be practically coincident at low frequencies yet far enough to encode the direction of arrival with a time lag between the microphones. The microphones also have an angle between them. In addition, since the distance between the microphones is usually approximately the distance between human ears, the near coincident approach can be understood as a rough approximation of the human hearing. A popular near coincident microphone pair is ORTF, where two cardioid microphones are placed $d = 17$ centimeters apart in an angle of 110^o, i.e., $\theta_1 = \theta_2 = 55^o$.

Fig. 3.5 [91] gives an summary of the widely used stereo microphone techniques.

Microphone Arrays

Stereo microphone setups clearly are not capable to capture the full 3D acoustic space. An obvious step towards multichannel microphone techniques is to extend

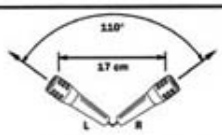
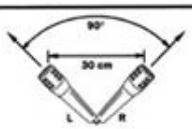
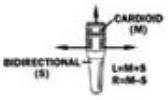
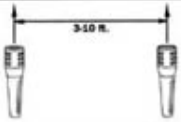
STEREO PICKUP SYSTEMS	MICROPHONE TYPES	MICROPHONE POSITIONS	
X-Y	2 - CARDIOID	AXES OF MAXIMUM RESPONSE AT 135° SPACING: COINCIDENT	
ORTF (FRENCH BROADCASTING ORGANIZATION)	2 - CARDIOID	AXES OF MAXIMUM RESPONSE AT 110° SPACING: NEAR-COINCIDENT (7 IN.)	
NOS (DUTCH BROADCASTING FOUNDATION)	2 - CARDIOID	AXES OF MAXIMUM RESPONSE AT 90° SPACING: NEAR-COINCIDENT (12 IN.)	
STEREOSONIC	2 - BIDIRECTIONAL	AXES OF MAXIMUM RESPONSE AT 90° SPACING: COINCIDENT	
MS (MID-SIDE)	1 - CARDIOID 1 - BIDIRECTIONAL	CARDIOID FORWARD-POINTED; BIDIRECTIONAL SIDE-POINTED; SPACING: COINCIDENT	
SPACED	2 - CARDIOID OR 2 - OMNIDIRECTIONAL	ANGLE AS DESIRED SPACING: 3-10 FT.	

Figure 3.5: Stereo Microphone Techniques

the techniques utilizing two microphones with additional microphones positioned in a desired setup. Various microphone array based techniques have been proposed to extend to capturing range as well as to increase spatial resolution, such as the linear array, the Decca tree, the spider array and the Fukada tree.

- **Linear Array**. In a linear microphone array usually omnidirectional microphones are positioned in a linear grid, either in a line or a plane, and they are physically pointing towards the sound source. This approach can be interpreted as a linear extension to an AB pair. Usually, the distance between two adjacent microphones in the grid is constant, thus making the analysis of the direction of arrival easier.

- **Decca Tree**. Three omnidirectional microphones are positioned in a tree-

like shape so that one microphone is physically pointing towards the sound source and the other two are physically pointing to the sides positioned 1.5 meters behind the front facing microphone. The distance between the side facing microphones is typically 2 meters.

- **Spider Microphone Array**. Five microphones are positioned at the ends of star shaped arms facing outwards from the center. The polar patterns of the microphones can vary, and some manufacturers provide systems with electrically controllable microphone polar patterns.

- **Fukada Tree**. This is similar to the Decca tree, but it contains seven cardioid microphones positioned so that three microphones form a triangle in a similar manner as in the Decca Tree. Two microphones are positioned the left and right side of the triangle, and the two remaining microphones are positioned to physically point to the back of the setup.

- **B-Format**. A B-format signal consists of four microphone signals, a pressure signal, a front-to-back signal, a side-to-side signal, and a up-to-down signal.

The possibilities are endless. Each microphone array setup has unique advantages and disadvantages. In our application, the microphone setup should provide sufficient number of microphones in order to generate accurate RIR measurements. It should also have been proven to offer good quality when used for recording musical signals so that we will have good references when comparing our results. Finally, it should be directly mappable to the mainstream speaker systems such as 5.1 systems. The spherical microphone array developed by Johnston et al [92] [93] meets all our requirements. This microphone array consists of 5 equal-angle spaced directional (hyper-cardioid) microphones in the horizontal plane, plus two highly directional (semi-shotgun) microphones aimed vertically up and down, spaced on a sphere of about sphere 30 cm size (0.9 milliseconds delay based on the speed of sound) [92], as shown in Fig. 3.6. Spherical microphone arrays have been proved to be effective for room acoustics measurement [94] . Subjective evaluation in [92] proved the effectiveness of this particular spherical array. It provides 7 simultaneous measurements from different angles when used to capture RIRs which enables robust analysis of the room responses. With minor adjustments of a 5.1 surround speaker setup, each of the speaker can be mapped directly to one of the 5 channels on the horizontal plane in this microphone array.

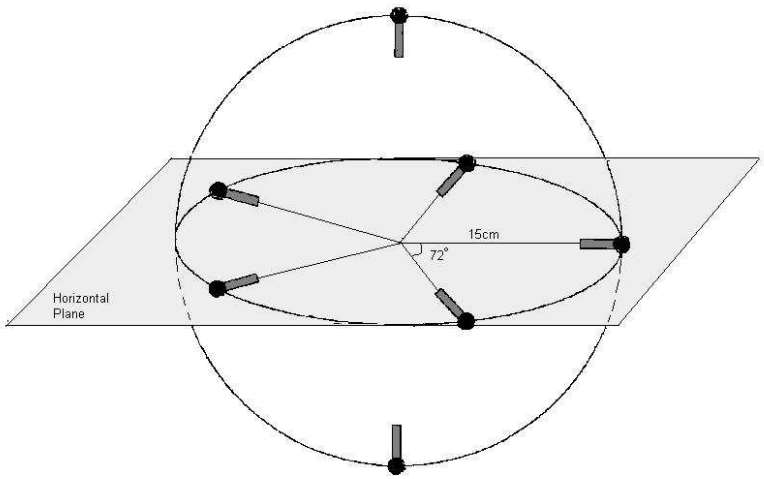


Figure 3.6: Microphone Array

3.1.3 Measurement System

Our measurement system works as follows. The chirp signal is generated by a laptop computer and played to a speaker. Assuming that most RIR would not exceed 3 seconds, we use chirp signals with a duration of 3 seconds and frequency sweeping from 0 to 24 kHz. At the receiver end, the output signals of a microphone array with 7 microphones are recorded to the same laptop through a multichannel audio interface, together with the unaltered chirp to be used as the reference signal. The unaltered loopback reference signal is important in that it eliminates the need for estimating the latency in the playback-record chain. To obtain the multichannel RIR, the received signals are correlated with the reference signal. Just as in a radar processing application, this function compresses the pulse and gives rise to the room impulse response that is to be analyzed [82].

The radiation pattern of the sound source (the speaker, in our case) must be considered in order to produce a set of truly representative impulse responses. Omnidirectional sources or sources with a wide radiation angle are generally preferred. This is because that highly directional sources are not able to generate the early reflections for all surfaces and may affect the quality of the measured reverb tails.

Experiments have shown that the chosen signal is able to capture the temporal and frequency characteristics, as well as to offer high SNR which is important in our case, i.e. to acquire room impulse responses (RIR's) for use in convolutions with dry

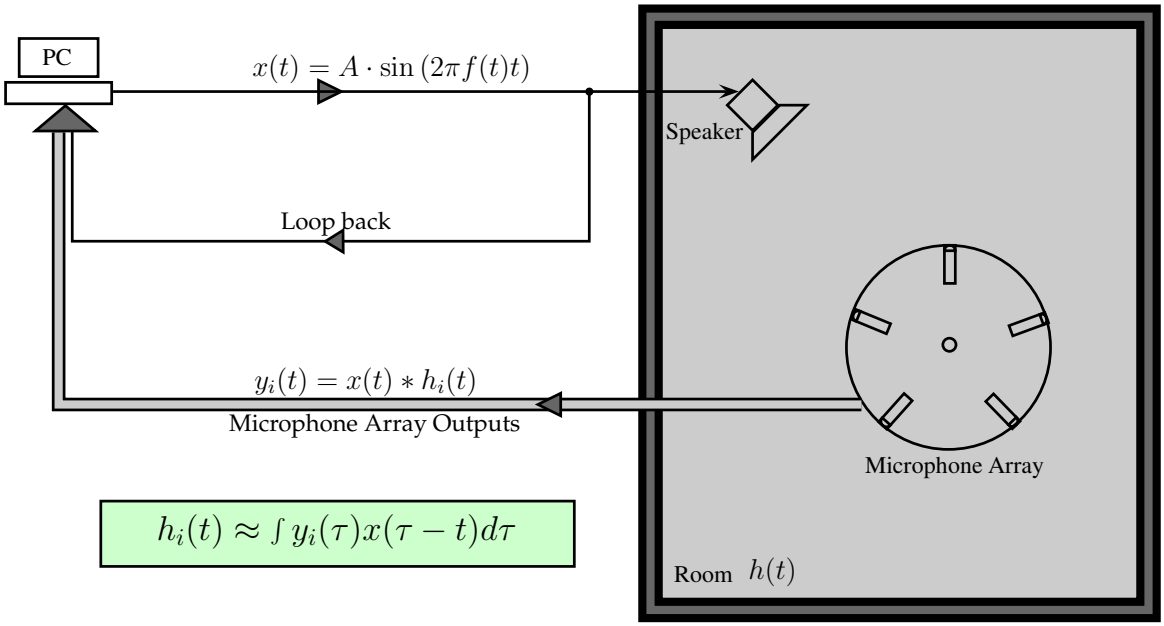


Figure 3.7: Measurement System

anechoic audio material. This is because any abnormalities in the reverberant tail of a RIR are easily recognizable due to the wide dynamic range of our auditory system and the logarithmic relationship between sound pressure level (SPL) and perceived loudness.

Sample output of of RIR measurement system is shown in Fig. 3.8 and Fig. 3.9 . It is worth mentioning that our method is independent of measuring techniques because what we need are the measurement results. In latter sections, we will discuss that a "good" analysis is the key to the success of our spatial sound rendering solution.

3.1.4 Equalizing Effect of Speaker/Mic Chain

The impulse response of the speaker-mic chain is measured by processing the sweep signal received at a microphone placed very close to the speaker. Because of the short distance between the speaker and microphone, the effect of absorption filters and reflections from the surfaces of the hall can be neglected. The the impulse response of the speaker-mic chain $h_{dry}(t)$ can be calculated using exactly the same procedure used in calculating impulse responses of the hall.

The measured RIR can be decomposed into the actual IR of the hall and the IR of speaker-mic chain $h_m(t) = h(t) * h_{dry}(t)$.

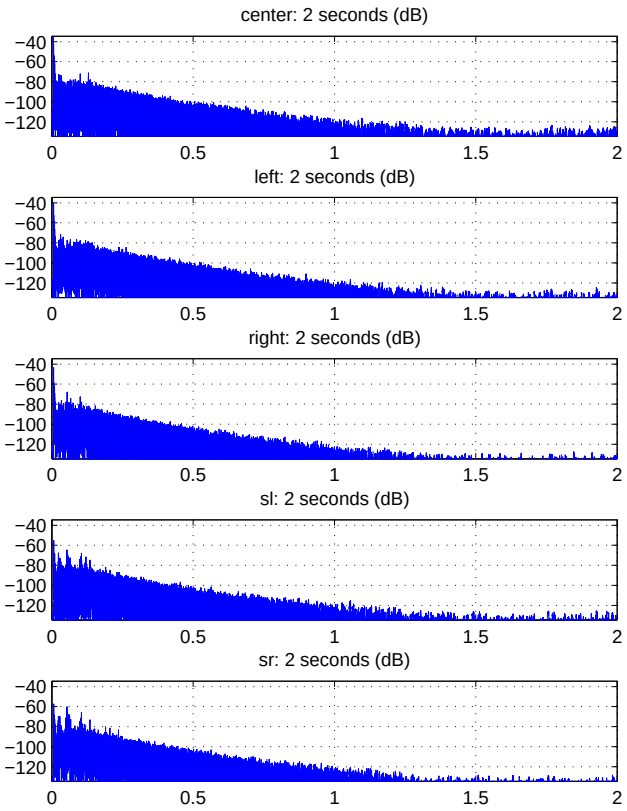


Figure 3.8: Typical Concert Hall MMRIR, Full Length

One way to cancel out the $h_{dry}(t)$ is to calculate its inverse filter $\hat{h}_{dry}(t)$ and to convolve with $h_m(t)$

$$\begin{aligned} h_m(t) * \hat{h}_{dry}(t) &= h(t) * h_{dry}(t)\hat{h}_{dry}(t) \\ &= h(t) \end{aligned} \tag{3.13}$$

Theoretically, this method is able to produce an accurate estimate of the actual IR $h(t)$. However, it is impractical in our case. This is because $h_{dry}(t)$ is generally a mixed phase filter which has zeros outside the unit circle,and as a result, its inverse filter $\hat{h}_{dry}(t)$ is not a stable filter with poles outside the unit circle. It can be converted to a minimum phase filter by computing its cepstrum and replacing any anticausal components with corresponding causal components [95]. In other words, the anti-

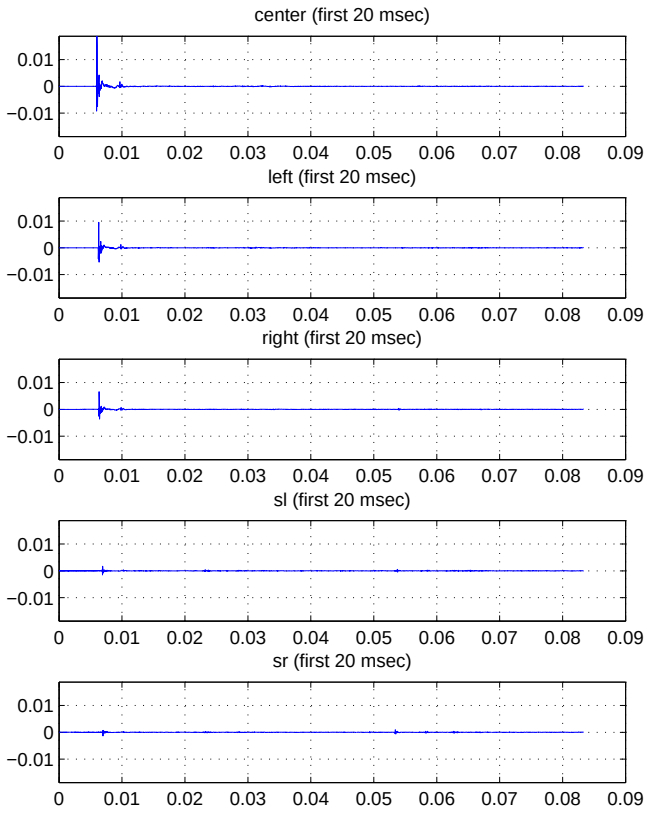


Figure 3.9: Typical Concert Hall MMRIR, First 20 milliseconds

causal part of the cepstrum, if any, is 'flipped' about time zero so that it adds to the causal part. Doing this corresponds to reflecting non-minimum phase zeros (and any unstable poles) inside the unit circle in a manner that preserves spectral magnitude. However, these methods modifies the phase characteristics of the filter heavily and may result in very unpleasant listening experiences. Given the fact that high end speakers and microphones have relatively flat frequency response in the region we are interested in, in the following discussion the effect of Speaker/Microphone on the measurements is ignored.

3.2 Analysis of the MMRIR

Various types of analysis can be performed upon the MMRIR to gain insight into the recording venue[84] [96], for example, Clarity C50 and C80, Centre Time, Early Decay Time, Reverberation Time RT30 and RT60, and etc. Because the purpose of our analysis is to build a image-source model, we focus on the analysis that leads to effective modeling of the wall and air absorption characteristics.

Completely modeling the effect of the air and reflective surfaces is extremely complicated, however previous studies [97] [59] [53] have shown that, for the purpose of spatial sound rendering, the effect of air and surfaces can be considered simply as frequency dependent absorption and such absorption can be modeled by low order IIR filters. The DIVA project proved the effectiveness of this simplification. Therefore, our analysis focuses on deriving a set of filters to represent the air and surface absorption. Unlike the DIVA system that uses the filters designed based on a model of the surface, the proposed method estimates an average absorption filter from the MMRIRs, which guarantees the consistency between the geometric model and the impulse responses.

3.2.1 Air Absorption Filters

The effect of air absorption is an important factor in image-source models, especially for large acoustic spaces, such as concert halls where higher order reflections can arrive considerably delayed from the direct sound. The absorption of sound in the transmitting medium (normally air) depends mainly on the distance, temperature, and humidity. There are various factors which participate in absorption of sound in air. In a typical environment the most important is the thermal relaxation.

Analytical Expression

The attenuation of sound in air as a function of temperature, humidity and distance have been studied and can be expressed analytically [98] as

$$g = 8.686 f^2 \times \left[\left(1.84 \times 10^{-11} \left(\frac{p_a}{p_r} \right)^{-1} \left(\frac{T}{T_0} \right)^{1/2} \right) + \left(\frac{T}{T_0} \right)^{-5/2} \times \left(0.01275 e^{-2239.1/T} (fr_o + \frac{f^2}{fr_o})^{-1} + 0.1068 e^{-3352.0/T} (fr_n + \frac{f^2}{fr_n})^{-1} \right) \right] \tag{3.14}$$

$$
\begin{aligned}
fr_o &= \frac{p_a}{p_r}\left(24 + 4.04 \times 10^4 \frac{0.02 + h}{0.391 + h}\right) \\
fr_n &= \frac{p_a}{p_r}\left(\frac{T}{T_0}\right)^{-1/2}\left(9 + 280he^{-4.170[(T/T_0)^{-1/3}-1]}\right)
\end{aligned}
\tag{3.15}
$$

where

- g is the air absorption factor, indicating the attenuation per meter in dB
- f is the frequency of the sound [Hz]
- p_a is the ambient sound pressure amplitude [kPa]
- p_r is the reference air pressure (101.325) [kPa]
- T is the absolute temperature of the air [^{o}K]
- T_0 is the reference air temperature (293.15) [^{o}K]
- fr_n is the relaxation frequencies for Nitrogen [Hz]
- fr_o is the relaxation frequencies for Oxygen [Hz]
- h is the molar concentration of water vapor (e.g. 0.4615 for 20% humidity)

The predominant phenomenon of the air absorption is observed as increased low-pass filtering as a function of distance from the sound source. For a given condition with temperature at 20^{o}C and humidity at 20%, (3.14) reduces to

$$
g = 8.686f^2\left(0.184 \times 10^{-10} + \frac{0.61424 \times 10^{-5}}{22842.0 + 0.43778 \times 10^{-4}f^2} + \frac{0.17393 \times 10^{-5}}{138.22 + 0.72348 \times 10^{-2}f^2}\right) \tag{3.16}
$$

This translates to a distance and frequency dependent attenuation as plotted in Fig. 3.10.

From a signal processing perspective, given a distance of d meters, the magnitude response of air absorption filter can be expressed as

$$
|H_d(f)| = 10^{\frac{-gd}{20}} \tag{3.17}
$$

and a standard filter design approach such as the Yule-Walker method can be used to design such a filter. Huopaniemi *et al* [97] also showed that first order IIR filters can be used to fit the resulting magnitude responses.

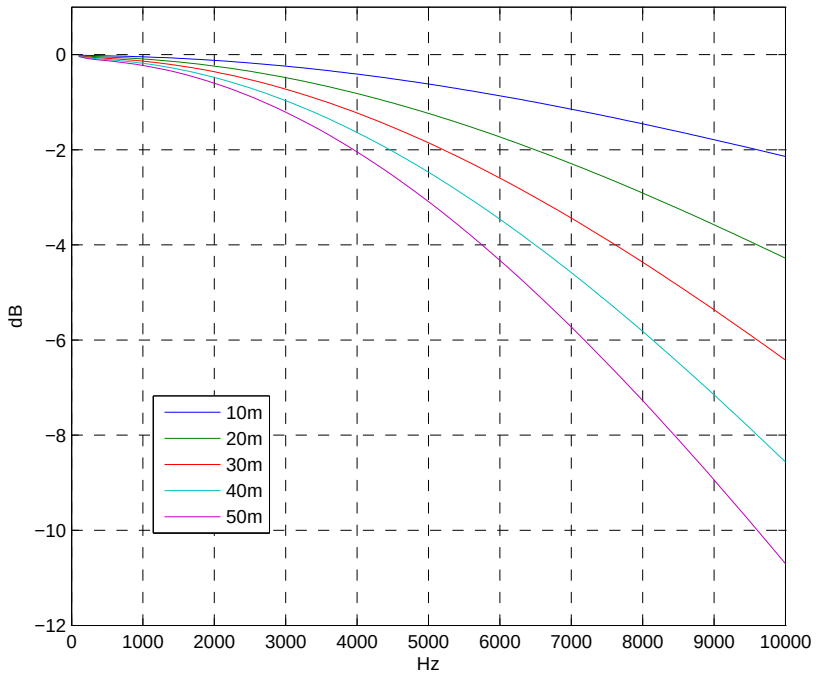


Figure 3.10: Distance and Frequency Dependent Air Absorption Attenuation

From Measured RIRs

The air absorption filters can also be estimated from the measured MMRIRs. Based on the assumption that the direct sound is only "filtered" by the air and attenuated by propagation, we can obtain the "impulse responses" of the air filter by retrieving the direct arrival (the first peak of a measured RIR) and its tail from a RIR. A set of direct arrivals with tails from RIRs at different distances but same angle are shown in Fig. 3.11. Now finding a IIR filter to represent the air absorption filter becomes a problem of an IIR filter with a prescribed time domain impulse response and can be solved using the Steiglitz-McBride algorithm [99]. Here we use second order IIR filters to approximate the air absorption filters and the resulting filters are shown in Fig. 3.11. Higher order IIR filters can be used to achieve better approximation if more accuracy is desired.

From measurements at different speaker locations, we can obtain a set of IIR filters for the corresponding distances. Fig. 3.11 shows direct arrivals in the RIRs at 3 and 5 meters and the resulting IIR approximations. The magnitude responses

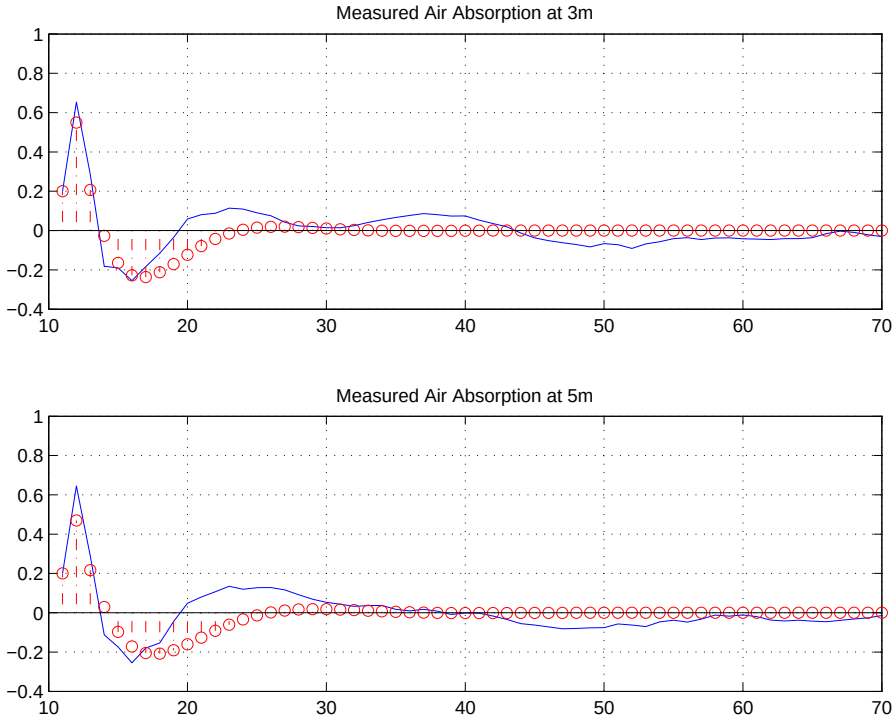


Figure 3.11: Measured Air Absorption Filter Impulses Responses and 2nd Order IIR Approximation

of the IIR approximation filters are shown in Fig. 3.12. It clearly shows the higher frequencies attenuate faster at greater distance. However, due to unknown errors occurred in measurements and inaccuracy of filter design process, the low frequency responses are not flat as one may expect by observing Fig. 3.10. This is an area for future investigation and improvement.

Parameterization

In many application areas, especially in the interactive applications, as the a sound source moves, the corresponding air absorption filter must be designed on the fly. Therefore it is desirable to parameterize this design process, which means to calculate the filter coefficients directly from the distance d. Lower order IIR filters generally are easier and more straightforward to parameterize. Therefore first-order IIR filters are used in the following parameterization process.

First we calculate the air absorption factor g_{5k} at 5kHz from (3.16)

$$g_{5k} = 0.0618$$

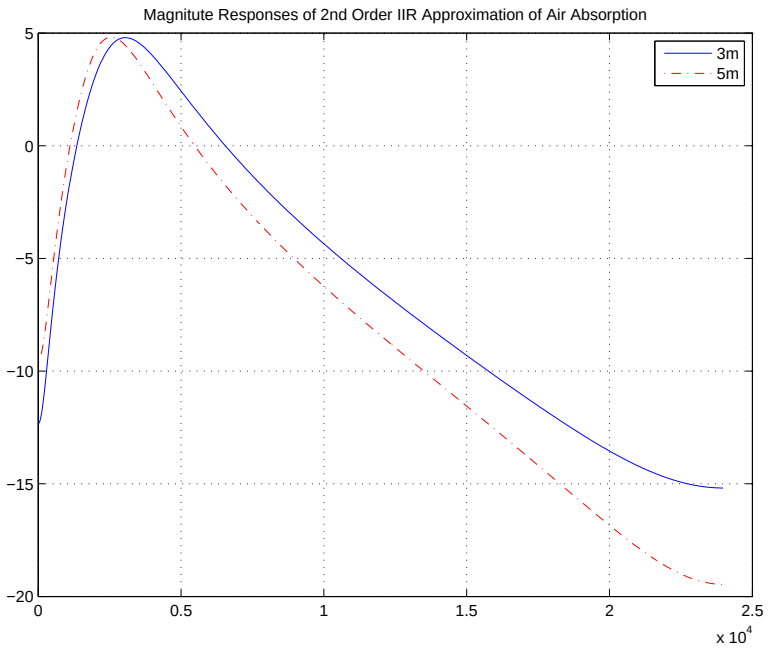


Figure 3.12: 2nd-order IIR approximation of air filter for different distances

and the the frequency response at 5kHz can be expressed as

$$|H_d(5kHz)|_{dB} = -g_{5k}d = -0.0618d \tag{3.18}$$

where d is the distance in meter. From $|H(5kHz)|$, we can determine the coefficient a of a first-order IIR filter represented as

$$H(z) = \frac{1}{1 + az^{-1}} \tag{3.19}$$

The relation between d and a is obtained by combining (3.18) and (3.19).

$$\left| \frac{1}{1 + 0.7934a - j0.6088a} \right| = \left(\frac{1}{1+a} \right) \times 10^{\frac{-0.0618d}{20}} \tag{3.20}$$

Expanding (3.21) gives an quadratic form of a as

$$(1 - p^2)a^2 + (2 - 1.5870p^2)a + (1 - p^2) = 0 \tag{3.21}$$

where $p = 10^{\frac{-0.0618d}{20}}$. Solving this equation results in two candidates for the value of a, and the one with a magnitude smaller than 1 is chosen to ensure the stability of the

IIR filter. Table. 3.1 and Fig. 3.13 shows the computed coefficients a and magnitude responses of the filters $H(z) = \frac{1}{1+az^{-1}}$.

distance	a
10.0000	-0.2233
20.0000	-0.3435
30.0000	-0.4254
40.0000	-0.4875
50.0000	-0.5373

Table 3.1: Coefficients of 1st-order parametric IIR approximation of air filters for different distances

Higher order IIR filters can also be used to approximate air absorption more precisely. However, the complexity of this parameterization process become forbidding and therefore it must be replaced by other types of filter design techniques such as Yule-Walker method [100] for IIR filters. Given the fact that a complex acoustic scene often involves thousands of sound sources that must be rendered simultaneously, higher order filters and the required design process are not practical choices.

Sometimes, it may be desirable to use FIR filters to approximate air absorption filters. This is because that on the modern computing hardware that has instruction pipelines and parallel processing units such as SIMD, FIR filters are easier to optimize and in turn more efficient than equivalent IIR filters. However, to achieve the same frequency responses, FIR filters generally require higher order and therefore more coefficients, which makes it impossible to parameterize and must be designed using other techniques such as the Parks-McClellan method [101].

In the image source approach described in the following chapters, we will use this parametric 1st order IIR filter to represent air absorption.

3.2.2 Wall Absorption Filters

The estimation of wall absorption filter is potentially much more complicated. First of all, the acoustic characteristics varies significantly across different types of surface materials. It is impossible to have a unified numerical representation. Table.3.2 and Fig.3.14 show the absorption characteristics of a few commonly used reflection materials.

Secondly, the temporal or spectral behavior of reflected sound as a function of

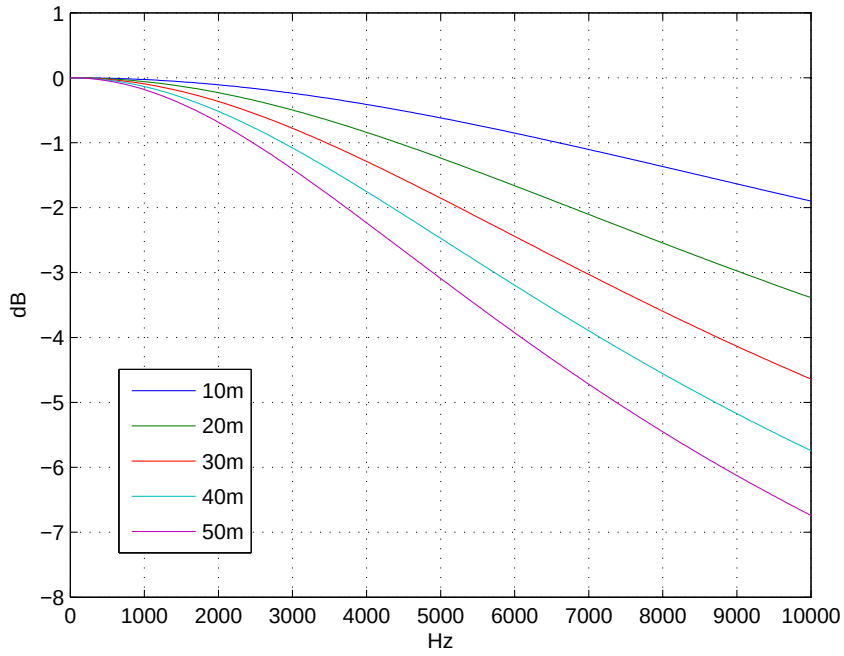


Figure 3.13: Magnitude Responses of 1st-order parametric IIR approximation of air filters for different distances

Materials	125Hz	250Hz	500Hz	1000Hz	2000Hz	4000Hz
smooth beton	0.01	0.01	0.01	0.02	0.02	0.02
gypsum	0.11	0.13	0.05	0.03	0.02	0.03
glass	0.30	0.15	0.10	0.05	0.03	0.02
wood	0.08	0.20	0.55	0.65	0.50	0.40
wool	0.15	0.70	0.60	0.60	0.85	0.90

Table 3.2: Frequency Dependent Absorption of Surface Materials

incident angle, the scattering and diffraction phenomena, etc., makes it impossible to use numerical models that are accurate in all aspects [59]. In our analysis, the goal is to obtain a filter to represent the *average* absorption characteristics of the surfaces in the target venue. With this in mind, we can then establish the frequency response of the wall filter from frequency-dependent Reverberation Time (RT60) based on the fact that the RT60 is almost solely determined by room dimension and wall material. According the the famous Sabine's formula [102], the reverberation time RT60 of an enclosure with volume V and boundary surface S, which is defined as the time it takes

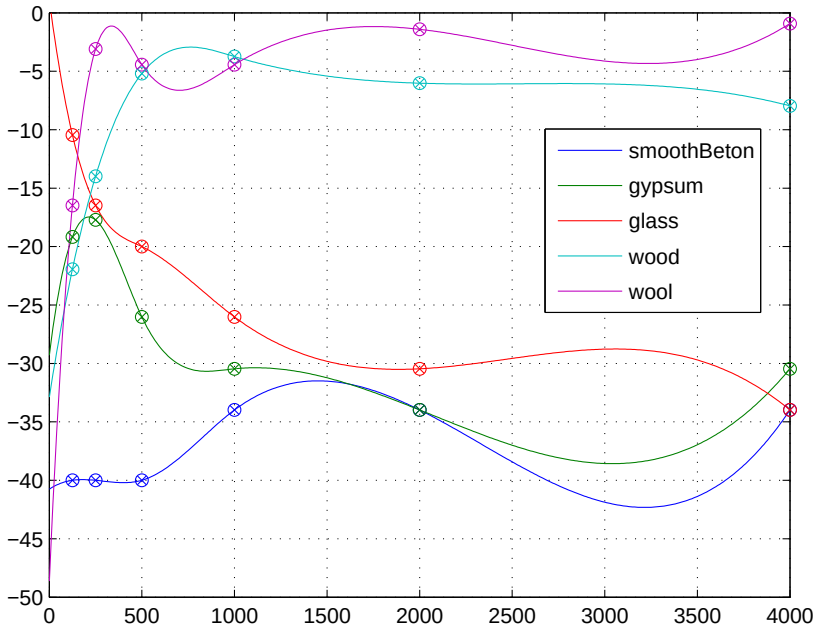


Figure 3.14: Frequency Dependent Absorption of Surface Materials

a signal to fall -60 dB, can be calculated by

$$RT60 = 0.163V/Sa \tag{3.22}$$

where a is the absorption coefficient averaged over the whole boundary. Because a is frequency dependent, RT60 is also frequency dependent. Our method is to estimate frequency dependent RT60 and then derive the wall absorption filter from it. RT60 can be estimated from measured RIR using various techniques, e.g., Schroeder's backward integration [102].

In our analysis, we decompose the RIR into a number of subband components using Short-Time Fourier Transform (STFT) with an FFT size of 2048, which gives us the frequency resolution of 23.44Hz (given the sampling frequency of 96 kHz) in each band. Short-time Fourier Transform of a typical concert hall RIR is shown in Fig. 3.15. The plot clearly indicates the decay rate is frequency dependent and generally high frequencies roll off faster than lower frequencies.

Then in each subband, an individual RT60 is estimated as the time it takes to decay to -60 dB of the direct arrival, as shown in Fig. 3.16.

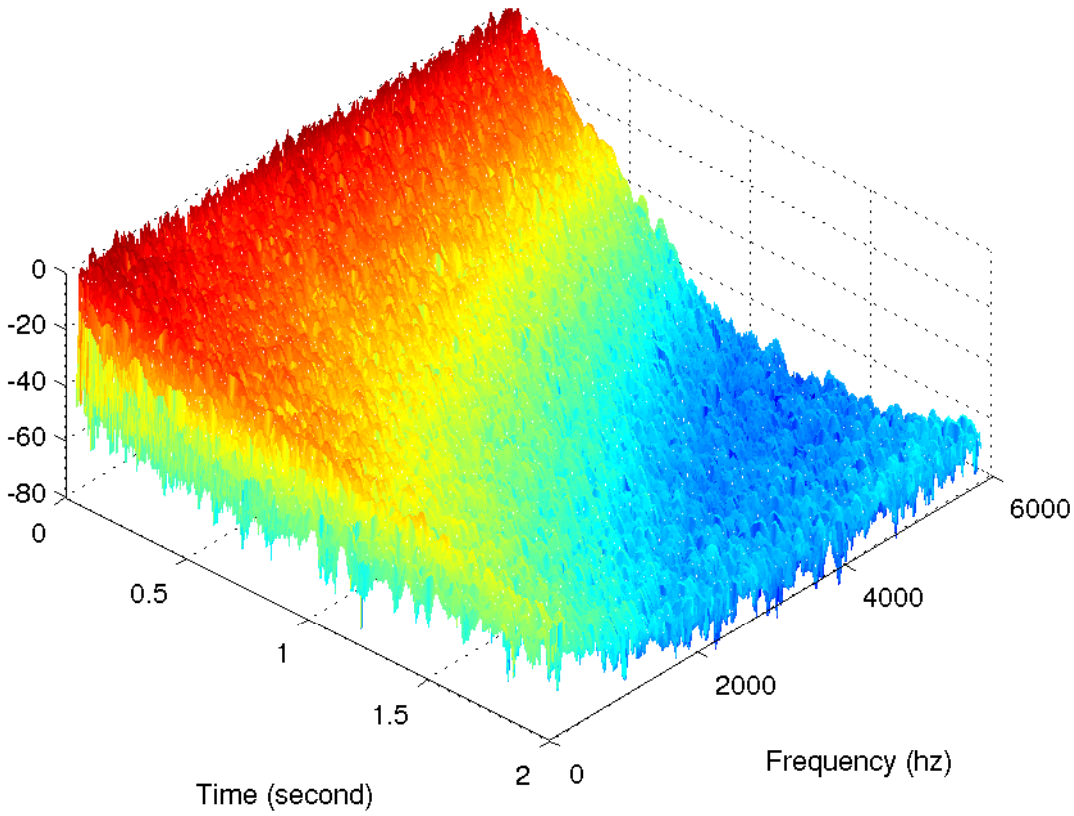


Figure 3.15: STFT of a Typical Concert Hall RIR

Having obtained the RT60, the next task is to estimate the frequency dependent wall absorption factors. Since the dimension of the room is know at the time of measurement (or can be estimated from RIR [103]) and RT60 indicates the time, and in turn the approximate distance d_{RT60} , that the sound has traveled before it reaches -60dB, we can estimate roughly how many times the sound hits the wall as

$$n = \frac{d_{RT60}}{dim_{average}} \tag{3.23}$$

The total wall attenuation in each frequency band is

$$w_{total} = \frac{-60dB}{p} \tag{3.24}$$

where $p = \frac{1}{d_{RT60}}$ is the propagation loss by $1/r$-law. Then the wall absorption in that

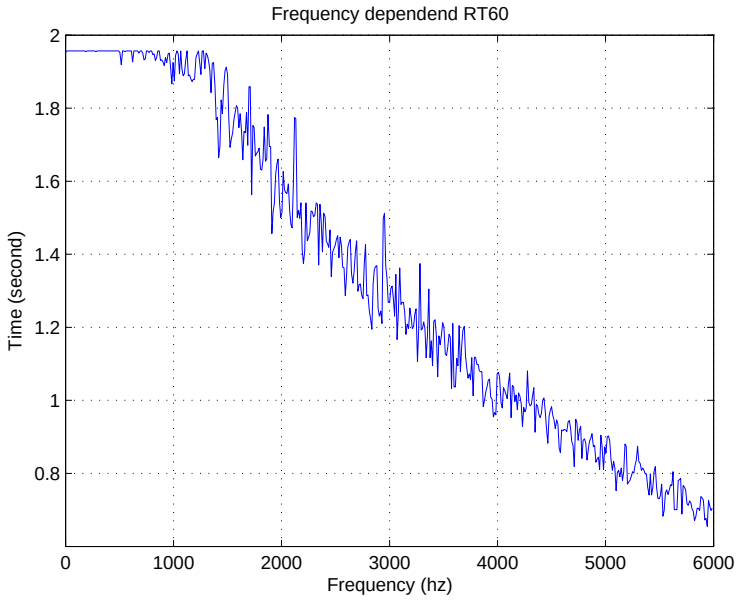


Figure 3.16: Frequency Dependent RT60

band is $w_{single} = w_{total}^{1/n}$. The frequency-dependent absorption factors composite the frequency response of the wall filter. Similar to the air filter, wall absorption can also be approximated by a second-order IIR filter [59]. Now the problem becomes designing a IIR filter from its frequency response and can be solved using standard filter design techniques. We use the Modified Yule-Walker Method [100] and the frequency response of the designed filter is shown in Fig. 3.17. The full band magnitude and phase responses are shown in Fig. 3.18.

The analysis can also be performed in a non-uniform frequency band, e.g. an auditory filterbank such as the gammatone filter bank, in order to make the analysis consistent with the human auditory system [104]. Note that because we have multichannel RIR measurements, the air filter and wall filter are estimated using all the channels. The target impulse response of the air filter is taken from the average of the normalized direct arrival tails. Similarly, the target frequency response of the wall filter is the average over all channels. The purpose of averaging over all available channels is to increase the signal-to-noise ratio. In the Section 4, we will describe how this approximation can be refined using our ray tracing model.

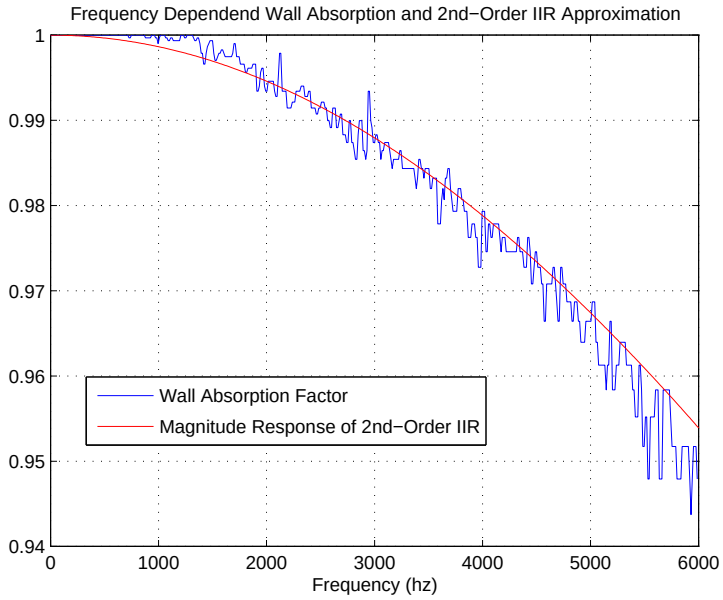


Figure 3.17: Frequency Dependent Wall Absorption and 2nd-Order IIR Approximation

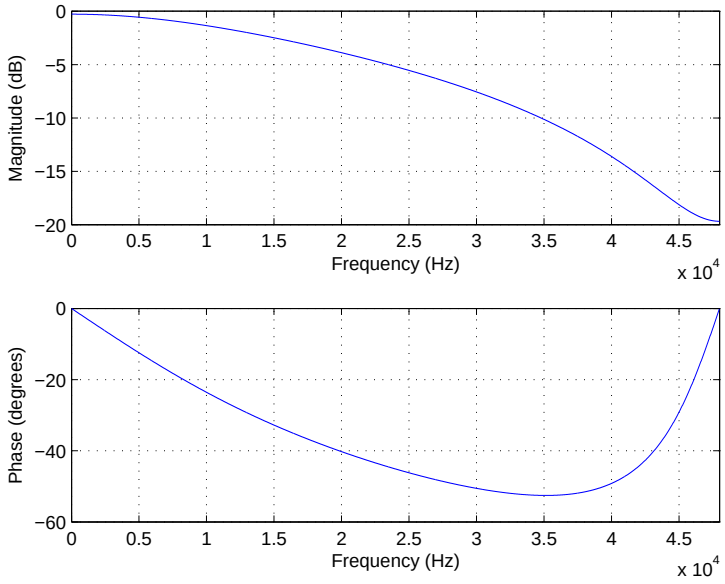


Figure 3.18: Frequency Responses of 2nd-order IIR Wall Filter

3.2.3 Reverberation Tails

One of the key assumptions made by the proposed hybrid approach is that the MMRIR acquired at one "good" position in the target space can be used to simulate the late reverb at other positions in the same space. The criterion for selecting a "good" measuring position include that this position must be representative and it must create a desired listening experience. For example, the middle of the audience seating area is generally considered a "good" position while the corners or the positions right next to the wall are not. The basis of this assumption is, as discussed in the previous chapters, that the late reverb does not change statistically (and perceptually) as the source or listener moves. The subjective evaluation proved the correctness of this assumption.

This assumption allows the hybrid approach to use segments of the MMRIR measured at one "good" position as the late reverb filters for rendering the late reverb at other positions in the same venue. Details of the rendering process are discussed in the next chapter.

Chapter 4

Spatial Sound Rendering

One of the most important tasks in building our spatial sound rendering system is to select an appropriate room acoustic model. This model needs to be scalable, easily controllable and able to render high quality at a reasonable complexity. Another important task is to customize this model using the analytic results from the previous chapter.

In our system where the real-time requirement imposes a limit on the computation complexity, we use a hybrid method that models only the direct sound and early reflections individually using the image-source method and simulates the late reverberation using a set of filters derived from the MMRIR.

Computational room acoustic modeling has been studied and used for more than three decades and a number of modeling schemes have been proposed. They can be largely categorized into wave-based methods, ray-based methods and statistical models [53]. Based on geometrical room acoustics, the ray-based methods are the most often used modeling techniques, while the other two types of methods do not fit into a real-time sound rendering system due to a number of reasons [53]. One of the most commonly used ray-based methods is the image-source method. The basic principle of the image-source method is to replace the reflected paths from the real source by direct paths from reflected mirror images of the source. In the image-source method the sound source is reflected at each surface to produce image-sources which represent the corresponding reflection paths. In our system, only a small number of early reflections are calculated with the image-source method due to its accuracy in finding reflection paths. Unlike the image-source models used in other auralization systems which need the user to specify the surface material characteristics [31], our image-source model uses what is derived from the MMRIR.

4.1 Image-Source Method

4.1.1 Finding Image Sources

The first step of image source method is to find all the required image sources based on the source position, listener position and room geometry. For a rectangle room, finding the image sources is straightforward and very efficient. In the case of a 2D room shown in Fig. 4.1, given the source position (x, y) and wall boundary coordinates (x_-, y_-) and (x_+, y_+), the first order image source (x', y') is simply

$$\begin{aligned} x' &= x + 2(x_+ - x) \\ y' &= y \end{aligned} \tag{4.1}$$

and second order image source (x'', y'') can be calculated from (x', y') by

$$\begin{aligned} x'' &= x' \\ y'' &= y' - 2(y_- - y') \end{aligned} \tag{4.2}$$

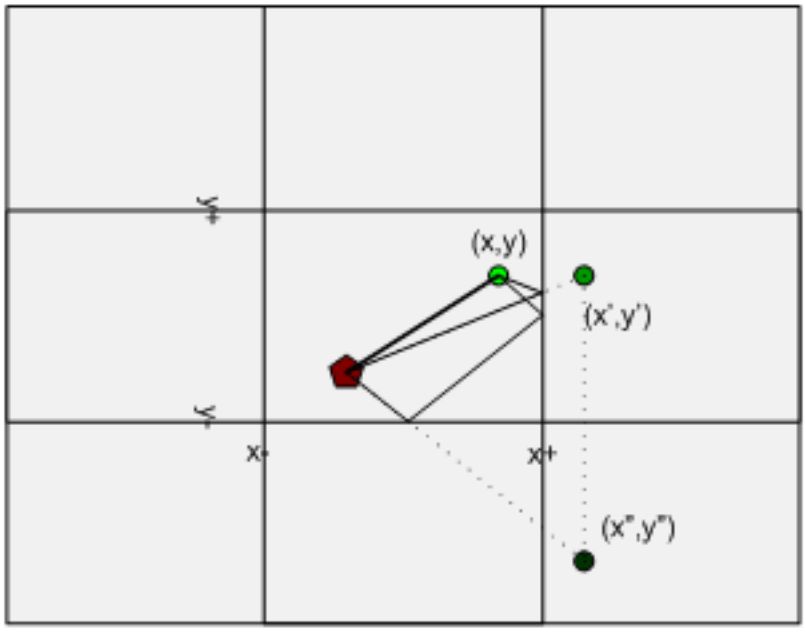


Figure 4.1: Finding Image Sources in a Rectangle Room

Higher order image sources can be found using similar methods as the image

sources of the nth order image source. These equations can be extended to 3D easily.

Finding the image sources for an arbitrary surface is much more complicated. First the surface must be (triangulated and) tested to see if the source and the listener are on the same side of the surface, then the following steps can be used to find the image source

1. The position of the mirror (image) point of the emitter off the surface is found;
2. The visibility of the mirror source is determined by the ray-plane-intersection detection algorithm [105]. The mirror point is visible if the intersect point $\vec{i}$ is inside the triangle specified by $\{\vec{v_0}, \vec{v_1}, \vec{v_2}\}$and inside the line segment from the mirror point to $\vec{p_l}$,
3. If the mirror source is visible to the listener, it is marked as a valid image source and associated with the reflectance properties of the surface.

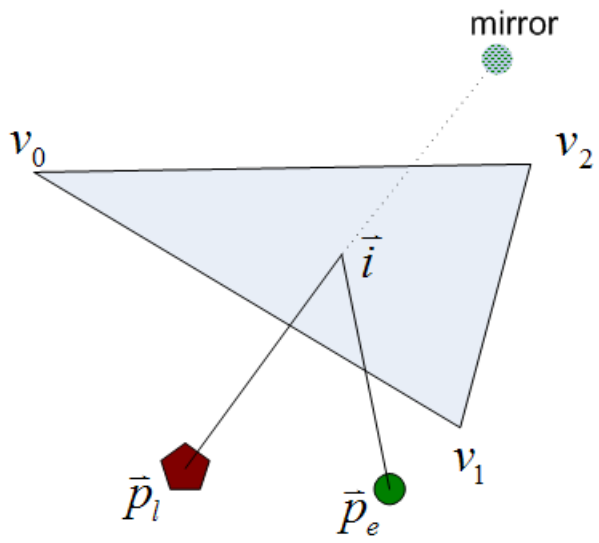


Figure 4.2: Finding Image Sources for an Arbitrary Surface

Despite the simplicity, the rectangle room based image source method works surprisingly well for many acoustic spaces that do not have regular shapes [106]. In this thesis, this method is used to generate all the required image sources.

4.1.2 Calculating Image Source Filters

For each sound source, these early reflections are modeled as a FIR filter which is called the early reflection filter $h_e(n)$ or $H_e(z)$ in this paper. If the air and wall absorption is ignored, this filter has a series of discrete peaks and each peak corresponds to the signal arrived from an image-source. When the effects of the air and wall are taken into account, each peak becomes a filter itself. In this thesis, this filter is refereed to as the image-source filter h_{is} or $H_{is}(z)$. Using the analysis result from the previous section, this filter can be expressed as

$$H_{is}(z) = H_{p,is}(z)H_{a,is}(z)(H_w(z))^{n_{is}} \tag{4.3}$$

where $H_{a,is}(z)$ and $H_w(z)$ are the air and wall filters obtained from MMRIR analysis, n_{is} denotes the number of times this image-source hits the wall (the order the reflections). $H_{p,is}(z)$ denotes the delay and attenuation from propagation derived from the distance between the (image) source and the listener by

$$H_{p,is}(z) = \left(\frac{d_r}{d}\right)^{\frac{1}{2}} z^{-(\frac{d}{c}f_s)} \tag{4.4}$$

where d is the distance between the source and the listener, d_r is the reference distance at which the input signal has a unit gain, c is the speed of sound and f_s is the sampling frequency. Note that $z^{-(\frac{d}{c}f_s)}$ may yield a fractional delay, and must be handled properly by resampling the audio signal. Various existing resampling techniques can be used, such as the linear interpolation, the Lagrange interpolation or the FIR based methods described in [107]. The fractional delay enabled geometric model allows for a seamless transition during the movement of the source or the listener that can not be achieved otherwise. The fractional delay also simulates the Doppler effect by changing the pitch of the audio signal according to the speed at which the source moves.

The reference distance d_r can be used to control the overall output level the image source method which is very useful if we need to adjust the level, for example, when trying to match the level of the early reflections and late reverberation.

An important assumption here is that the surface material must be homogeneous, meaning that the absorption characteristics are the same across the entire space. This is often not sure and requires individual modeling and processing for each image

source. However, this is not necessary in our approach because, as mentioned in the previous chapter, the wall absorption filter obtained from MMRIR analysis is an average filter that represent average absorption characteristics in the target space.

Because the signals arriving at the receiver (microphone array) are the superposition of direct arrival (actual sound source) and all the reflected copies (image sources), we can express the early reflection filter as

$$H_e(z) = \sum_{is} H_{is}(z) \tag{4.5}$$

Here we consider the actual sound source as a image source that has been reflected 0 times so that we have a unified representation.

4.1.3 Randomization

The image source method by itself does not consider whether the phase of the reflected sound is inverted or not. Therefore the generated impulse response has only positive amplitude. In real world, the inversion of phase is largely determined by the incident angle. However, since human ear is not sensitive to the phase information of an individual reflection, we do not need to accurately model the phase inversion. Instead, a randomized phase inversion can be implemented efficiently without sacrificing the realism.

One of the disadvantages of ray-tracing based methods is that the wall is often supposed to be perfectly flat and have constant absorption characteristics everywhere. In order to add the impression of diffused reflection, we impose randomness on the reflection angles by adding a small gaussianly distributed random number to the calculated position of image-sources,

$$\vec{x}' = (1 + \beta_x)^n \vec{x} \tag{4.6}$$

where $\beta_x \sim n(0, v_x)$ and n is the order of reflection. The roughness of the reflecting surface can be easily controlled by the variance v_x. We do the same to the wall absorption factors based on the assumption that the material on the reflecting surface is uneven to a certain degree. For simplicity, we use a universal random factor for the entire frequency range. The modified image-source filter becomes

$$H'_w(z) = (1 + \beta_w) H_w(z) \tag{4.7}$$

where $\beta_w \sim n(0, v_w)$. Similarly the unevenness can be controlled by the variance v_w.

4.1.4 Practical Consideration

However, the above mentioned method alone does not fit into a real-time framework because the number of image-sources grows exponentially as a function of the order of reflections, and it is computationally inefficient to use the image-source method to find the higher order reflections. Also, the inaccuracy in the estimation of air and wall absorption filters will propagate and accumulate as the order of reflection grows, which will potentially result in a synthesized reverberation that differs significantly with the real one. In other words, the image-source method is not a practical choice for simulating the late reverberation.

Fig. 4.3 shows the resulted impulse response of the image source method with 1st order image sources.

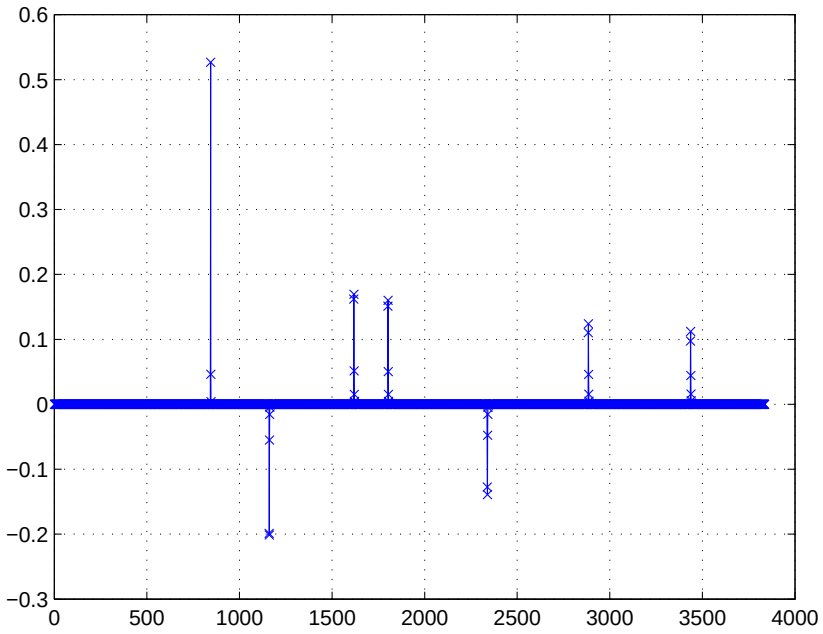


Figure 4.3: Result of the Image Source Method

4.2 Adding Reverberation

The late reverberation in a room is often considered nearly diffuse and the corresponding impulse response exponentially decaying random noise [30]. Under this assumption, the late reverberation does not have to be modeled individually for each source or listener location because it does not contain information for critical directional perception. To optimize computation in late reverberation modeling, a number of artificial reverberation algorithms have been proposed, e.g., [28] [31]. However, these algorithms are derived from (often overly) simplified physical model or perceptual models that are not fully established. Therefore they are often incapable of creating the acoustic impression of a sound space faithfully. Additionally, all these artificial reverberators contain multichannel feedback network so that the stability is not always guaranteed, especially when tuning the parameters. With the RIR measurement at hand, we have the power of re-creating the actual acoustic impression of the recording venue. One straightforward way of generating multichannel late reverberation is to convolve the dry signal with the tails of MMRIR directly. This method has the advantage of preserving the exact acoustic field at locations where MMRIR is made. Together with the early reflections generated by the model built upon the actual impulse response of the same recording venue, the consistency between the early reflections and the late reverberation, and the consistency between the synthesized impulse responses and the real ones are guaranteed.

Another very important characteristic of reverberation is the correlation of the signals that reach your ears. In order to give a listener a real feeling of the 'spaciousness' of a big room, the sounds at each ear should be somewhat incoherent [9]. This is why single channel measured RIRs are not capable of creating this incoherency. It is possible to artificially generating the incoherency by, for example, randomizing the phase with all pass filters. However, doing this inevitably modifies the characteristics of the RIRs. In our system, we have multichannel RIRs and each channel is inherently incoherent. Therefore it is a straightforward choice to use one channel of RIRs for each channel of outputs. The solution that uses single channel RIRs with decorrelating all pass filters may also be useful when the efficiency is at a higher priority that the spatial quality. Depending on the decorrelation techniques, the quality of the rendered late reverb may degrade considerably when compared to the quality using the true multichannel reverb tails.

4.2.1 Merging Early Reflections and Late Reverberation

Because in our system, the early reflections and the late reverberation are generated from very different approaches, how to merge the reverberation tail with early reflections become crucial [78]. We identify three aspects that need to be address in order produce convincing results.

1. The early reflections and late reverberation must reflect the same frequency characteristics. This has been address by using the absorption filters derived from the MMRIR to build the image source model.

2. They must be combined in a way that can cover the entire IR time span seamlessly.

3. Their levels must match.

Method 1: Crossfading

The straightforward approach is to cross-fade between the impulse responses of the early reflection filters derived from our image source model and the reverberation tails, with a predetermined cross-fade curve and predetermined cross-fade point. A linear cross-fade curve is shown in Fig. 4.4.

The crossfade point can be variable depending on the largest delay of the reflections derived from the image source model. For simplicity and efficiency reason, the crossfade points in the current implementation are static, meaning that these points remain constant throughout all the sources.

Method 2: Replacing The Real Reflections with The Synthesized

One alternative to the crossfading method is to replace the reflections in the recorded RIRs with the ones obtained from the image source model. This involves identifying individual reflections in the recorded RIR and insert zeros to replace these reflections and the trailing tails. With this method, there is no need to find a cross-fading point and the RIR with early reflections removed can be used directly as the late reverberation filter. The disadvantages are that the reflections may have to be removed manually.

This process is illustrated in Fig. 4.5, where the red circles enclose the individual early reflections need to be removed. Once the actual reflections are removed, the synthetic early reflection filter can be simply superposed on top.

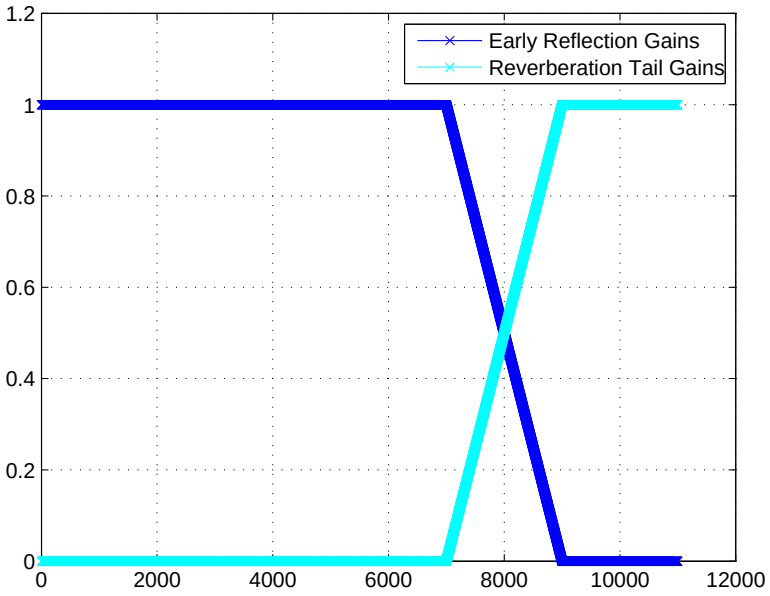


Figure 4.4: Result of the Image Source Method

Preliminary evaluation shows, if properly implemented, both the cross-fade method and the reflection replacing method are able to producing convincing results.

Matching The Levels

Matching the levels between the early reflections and late reverberation is another critical requirement. The relative level greatly affect the perception of spatial depth and overall spatial impression and it also affect the timbre of the sound (see Table. 1.1). Two approaches of matching the levels are developed in this thesis. It worth noting that both approaches uses the reference distance d_r for tuning the image source model.

By Matching the Reverberation. This approach is based on generating late reverberation from the image source model and then adjusting the image source model to obtain a matching reverberation tail to the recorded RIR. It contains the following steps,

1. Analyze IR, calculate a few key parameters (RT60, RT20, RT30...) and generate preset including air and wall absorption filters;

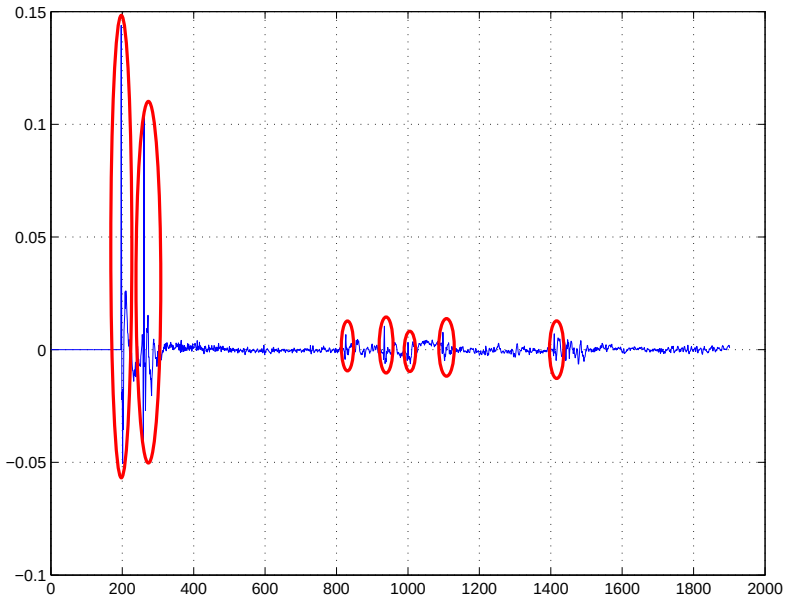


Figure 4.5: Replacing The Real Reflections with The Synthesized

2. Run high order ray tracing with an impulse as excitation signal to generate IR;

3. Calculate a few key parameters (RT60, RT20, RT30...) from the synthesized IR;

4. Compare these parameters with those from real IR, adjust the filters and d_r based on the differences *manually*;

5. Repeat 2-4 until the differences are below certain level.

While this approach is very accurate in terms of find a d_r that leads to matching these two parts and there is no need of knowing the recording setup, it is an iterative approach that is extremely time consuming.

By Matching the Direct Arrival One efficient alternative is to derive the direct arrival only from the image source model based on the source/microphone locations where the RIR is recorded, then adjust d_r so that the level of the synthetic direct arrival matches the recorded one. It contains the following steps,

1. Calculate direct arrival levels using the exactly recording configuration in ray tracing.

2. Compare with the direct arrival levels of real IR (first peaks);

3. Calculate d_r from the difference between the level of the synthetic direct arrival and recorded direct arrival.

The accuracy of this approach relies on an accurate air absorption, the quality (in term of noise and distortion) of the RIR, and the knowledge of the RIR recording setup. In most cases, it is reasonable to assume that the source and the microphone locations are known at the time of making the measurement and can be recorded as part of the measurement and that the source and the microphone are not very far apart so that the impact of the air absorption can be ignored. Therefore, this method is able produce good d_r estimation with minimal effort.

In this chapter, the details of our spatial rendering system have been laid out. Next, we will discuss the development of a real-time implementation of this system.

Chapter 5

Implementation and Evaluation

5.1 Implementation

Our spatial sound rendering system is implemented as in Fig. 5.1. It consists of two main components, namely, online processing unit on the left and offline processing unit on the right. There is also a control unit that controls the analysis and rendering process.

5.1.1 Offline Unit

The offline processing unit handles the tasks that do not need to be done in real-time, including preprocessing and analysis of RIR and extracting the reverb tails. The offline unit needs to run only once when a new set of RIR measurement is fed in or critical control parameters such as the orders of the absorption filters need to be adjusted. The analysis results from a certain set of RIR, including absorption filters, reverb tail and optionally room dimension and geometry, are grouped together as a "PRESET". A preset may contain several "profiles" that are targeted for different complexity and real-time requirements. Creating profiles is fairly straightforward by controlling, for example, the order of the absorption filters and the length of the reverb tail. The supplier who provides new presets can keep its offline unit and raw RIR data from the users by shipping the preset data only. In this way, the supplier is able to improve their measurement and analysis without affecting the end users. The offline unit may also contain an optional preprocessing block that is responsible for outputting "nice and clean" RIR's by, e.g., normalizing, removing the distortion caused by the playback-recording chain and an inappropriate source signal, and/or

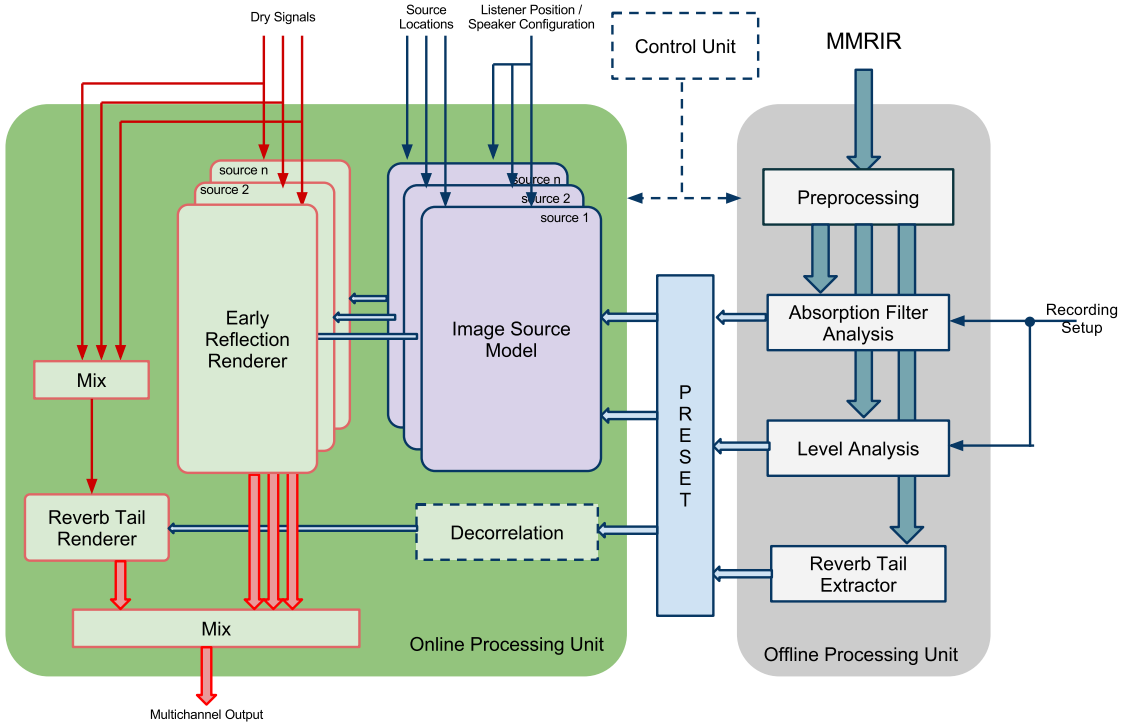


Figure 5.1: System Diagram

removing noises. In order to refine the estimated absorption filters, a high order image-source model can be used to synthesize the MMRIR and absorption filter can be adjusted iteratively so that the synthetic MMRIR matches with the measurements statistically or perceptually.

5.1.2 Online Unit

The online processing unit is responsible for rendering the "dry signals" over a multichannel speaker system in such a way that the perceived sound sources are located at the user determined positions in the recording venue. The online unit contains an image-source model for generating the early reflection filters based on the room geometry, speaker configuration and the user-defined source and listener positions in real-time, a signal processing chain consists of an early reflection renderer and an reverb tail renderer, and optionally, a decorrelator and a preprocessor. The decorrelator may become necessary when using the same set of reverb tails to render multiple sources over multiple speakers.

The image source model finds all the (virtual) sources and calculates information

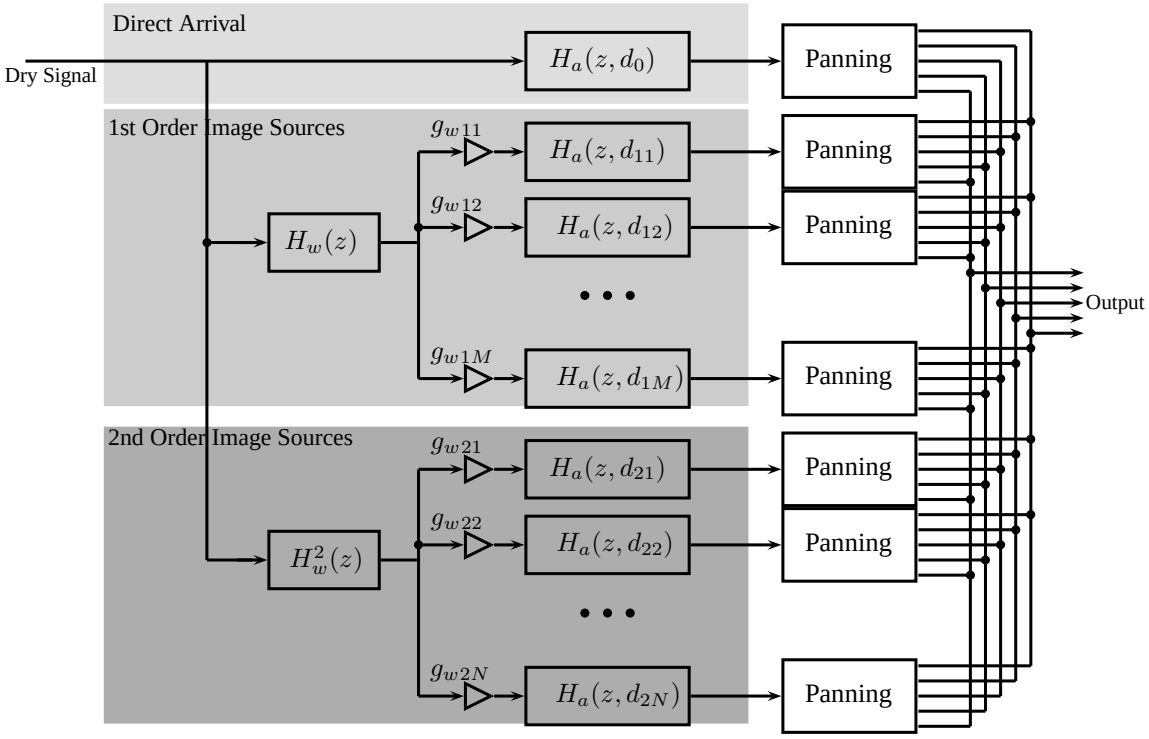


Figure 5.2: Image Source Renderer

required by the early reflection renderer for the first few orders of reflections using the methods developed in the previous chapters. These include coefficients for air absorption filters $H_a(z, d)$ and wall absorption filters $H_w^n(z)$, the wall randomization factor g_w and coefficients for the panning block.

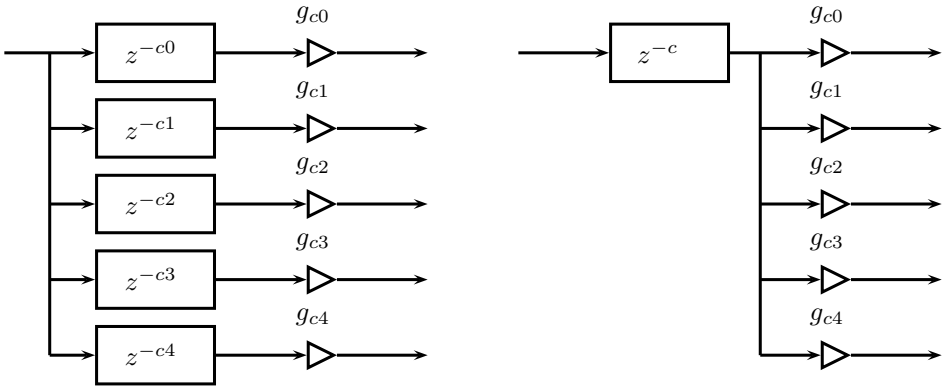
The contents of the panning coefficients depend on what panning techniques are used. In the case of VBAP, the coefficients include the one gain for each output channel where two of the gains are non-zero. While for the physically accurate panning, we need one delay and one gain for each output channel.

Image source model only needs to run when there is any change in the input information, namely the room geometry, speaker configuration or the source and listener positions.

Fig. 5.2 depicts the diagram of the early reflection renderer. The virtual sources are grouped together by the order of reflections. The actual source passes through an air absorption filter $H_a(z, d_0)$ before being fed into the panning block. To render the group of Nth order image sources, cascaded N wall absorption filters $H_w^N(z)$ are applied to the dry signal, and then each image source in this group is processed with

a wall randomization factor, an air absorption filter and a panning block. Note that each air absorption filer has a unique set of coefficients that are calculated by the image source model based on the distance.

The duty of the panning block is to "upmix" the mono dry signal to multichannel signals for the target speaker system configuration. For example, for a 5.1 system, the panning block generate 5 signals from the signal channel input. Fig. 5.3 shows the panning block used in the current implementation, as well as an VBAP based panning block. The outputs of all the panning blocks are mixed together to form a signal multichannel output as a final result of the early reflection renderer.



(a) Full Panning (b) VBAP Panning

Figure 5.3: Panning Block

To render the late reverberation, all the dry signals are first mixed into one mono signal and then fed into a FFT filter block that contains 5 FFT filters, each one corresponds to one output channel. The details of the FFT filtering can be find in many textbooks [108] and will not be repeated here.

The FFT filter block converts the mono mixed dry signal to a multichannel signal that only contains the rendered late reverberation. Once mixed with the result of the early reflection renderer, a convincing multichannel spatial sound rendering is completed.

The above mentioned system is implemented on PC to demonstrate the online processing unit. The high efficiency is proven by the fact that when rendering one sound source, the program only uses small percentage of the CPU resource on a modest computer with a dual core 2GHz AMD Athlon X2 CPU [1]

[1]This program only utilizes one core.

Offline unit

MMRIR	7-channel, sampled at 96kHz, effective length of 2.5 second, recorded at Rolston Hall of the Banff Center
Preprocessing	Normalization
Air Absorption	1st-order IIR
Wall Absorption	2nd-order IIR
Level Matching	Match the direct sound of a known recording location
Tail Extractor	retain reverb tail from 0.1 sec to 2 sec, resampled to the working frequency of online unit, 44.1 kHz

Online unit

Source listener	a number of source positions were test, listener (microphone array) sits on the middle point of the room. Source and listener are assumed omni-directional. Single static source.
Image-source model	Shoebox geometry is used with the estimated dimension of Rolston Hall (22m x 17m x 6m); order of reflection is 1; variance of the position and the wall randomization factor are both 0.1.
Dry Sound	clarinet recording at 44.1 kHz
Reflection filter	Delay-filter-sum
Reverb filter	Fast convolution using FFT
Panning	Delay and amplitude panning

Table 5.1: System Configuration and Preset

5.1.3 Example Configuration

After experimenting with various parameters, a system configuration is selected as in Tab. 5.1 to build a testing system that is capable of offering good quality at a reasonable complexity to run in real-time. This testing system is an example of a "home production" profile that is targeted for home based music creation that requires a balanced quality and complexity.

5.2 Subjective Evaluation

Human auditory perception is extremely complex so that objective evaluation methods based on modeling the auditory system are not capable to produce truly meaningful assessment of the audio quality which is very subjective in nature. Therefore, although a number of objective evaluation schemes are available for various applications [109], in this thesis, I choose to primarily rely on user study and subjective assessment to determine the quality of audio processing systems. Objective evaluation can also be used as an complementary source for quality assurance.

The goal of the proposed system is to have the ability of placing and moving sound sources in the target venues freely and meanwhile retain the spatial impression faithfully. Therefore the subjective evaluation focuses on

1. to determine whether the proposed system is able to duplicate the spatial impression closer than other existing solutions;

2. to determine how the various parameters and implementation techniques affects the resulting quality.

One question left to answer is accuracy of sound source placements. This is greatly affected by the panning techniques being applied and is not included in the evaluation in this thesis. However, preliminary evaluation results show that the accuracy is comparable to, if not better than, the existing solutions.

5.2.1 Methodologies

Spatial audio reproduction is a relatively young topic and evaluation methodology is yet to be standardized. Although researchers have proposed several schemes [110] [111], no widely accepted methodology to study auralization quality or quality of spatial sound reproduction is agreed upon within the research community. Therefore, this thesis uses the standard evaluation methodologies that are originally designed for assessing quality of audio codecs which include multi-channel audio encoder/decoder.

MUSHRA

MUSHRA stands for Multi Stimulus test with Hidden Reference and Anchors and is a subjective evaluation method specified in [112]. This test method was developed in

the late 1990's when Internet bandwidth became large enough for audio streaming. It was officially released in June 2001.

MUSHRA uses a reference (the original signal) and one or more anchors (processed original signal). The reference is necessary for grading the annoyances of the various artefacts. Assessors are asked to judge their degree of similarity of each anchor signal to the reference signal. The goal is to find the most faithful algorithm. The assessor can freely switch between all test signals, the reference and the anchor(s). The possibility of direct comparison between systems gives a high degree of resolution.

A-B Comparison

A-B Comparison is a preference evaluation. Each assessor is given two signals and is asked to choose which one he or she prefer. The goal is to find the more pleasant sounding algorithm.

5.2.2 Environment and Procedure

The evaluation environment was setup in a medium size laboratory with a 5.1 surround speaker system. This environment does not the offer the acoustic properties for an interference free listening experience, such as very absorbent interior surface, noise isolation and high quality speaker system. However, to some degree, it resembles the common living room environment which is one of the main target application areas of the proposed system. With modest confidence, the subjective approval in this environment proves the effectiveness of the proposed system for the common living room environment.

The subject group consists of undergraduate students, graduate students, faculty and staff who have adequate knowledge in the domain of digital audio and sound production. 21 participants have completed the procedure. The reference reproduction level is set according to estimated average age of this group (25 years old) according to the principles outlined in [113].

Each participant was asked to complete a procedure that consists of a training session, a similarity (MUSHRA) trail session and a preference (A-B comparison) trail session. The purpose of the training session is to familiarize the participants with the interface and the procedure, as well as to establish a valid scoring range between the "best" and the "worst". In order to evaluate various aspect of the proposed framework, the MUSHRA session and the A-B comparison session contains

9 similarity trails and 6 preference trails respectively.

In the MUSHRA session, each trail contains one reference signal and four anchor signals (some trails have three anchors and one hidden reference). The assessor is asked to rate the similarity of the four anchor signals to the reference signal on a scale of 0 to 100.

In the A-B session, each trail contains two signals. The assessor is asked to rate his or her preference on a scale of 0 to 3.

The complete lists of the trails and procedures are detailed in Appendix. A.1.

5.2.3 Results and Analysis

For each set of scores, the highest one and lowest one are removed from the set. Then an average score and its deviation is calculated from the remaining set.

Comparing to the Artificial Reverberation

The artificial reverberation used in the evaluation include 4 of the reverberators provided by Adobe Audition™2.0, namely "Full Reverb", "Studio Reverb", "Reverb" and "QuickVerb" [2]. The first three are targeted towards high quality music production while the fourth one is an efficient alternative. All of the artificial reverberators have a set of adjustable parameters that include room size and reverberation time, among others. To establish valid comparisons, all the artificial reverberators are tweaked to have the same room size and reverberation time. Fig. 5.4 shows that the proposed hybrid system offers a much closer resemblance to the actual recording. Fig. 5.5 shows that it creates a much more pleasant listening experience for the listeners.

Various Aspects within the Proposed Hybrid System

Another focus of the subjective evaluation is to study the impact of various parameters and implementation techniques on the rendering quality.

Effect of Reverberation Tail Length The effect of the length of the reverberation tail is studied. As expected, this depends on the nature of the target room. To simulate a more reverberant room requires long tails than a drier room, as shown in Fig. 5.6.

[2] "QuickVerb" appears to be a mono reverb, therefore the comparisons with "QuickVerb" should be ignored.

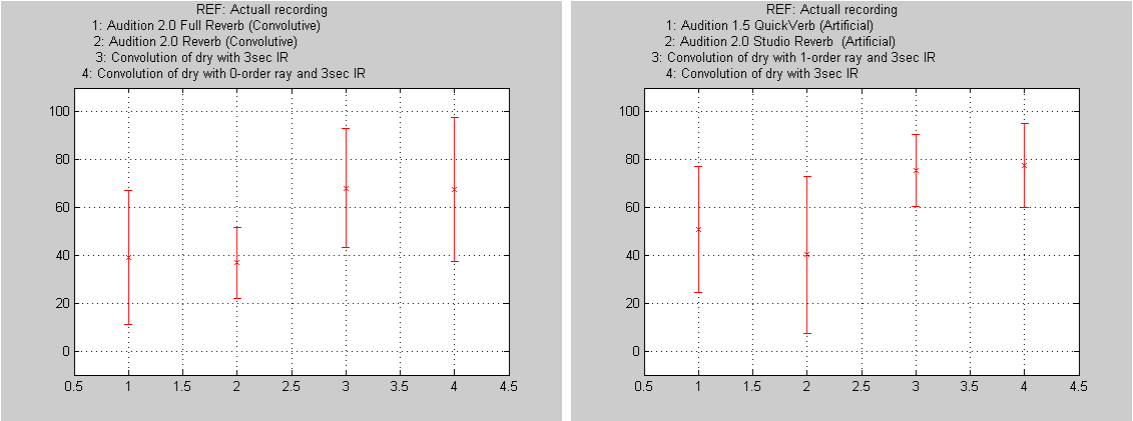


Figure 5.4: Comparing to the Artificial Reverberation - MUSHRA

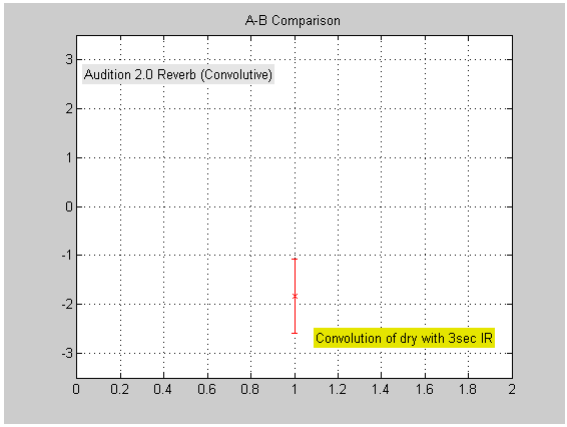


Figure 5.5: Comparing to the Artificial Reverberation - A-B

Effect of Order of Reflections The result in Fig. 5.7 suggests that the order of reflections used in the early reflection renderer does not affect the quality greatly.

Mismatching Late Reverberation The result in Fig. 5.8 shows the results of using the reverberation tails from other venues with the correct early reflections to reproduce the target venue. It clearly states that reverberation tail is the deciding factor of retaining spatial impression.

Mismatching Early Reflections The left figure in Fig. 5.9 suggests that to synthesize the surround recording at position A, the direct arrival from A along with the reverberation tail from position B is not sufficient. This is not surprising since beyond the direct arrival, first few reflections also carries important information about the

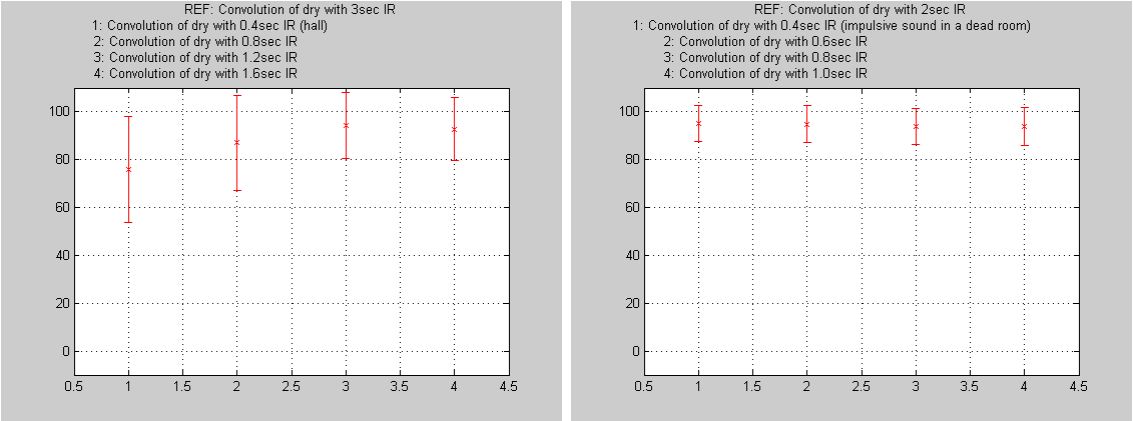


Figure 5.6: Effect of Reverberation Tail Length

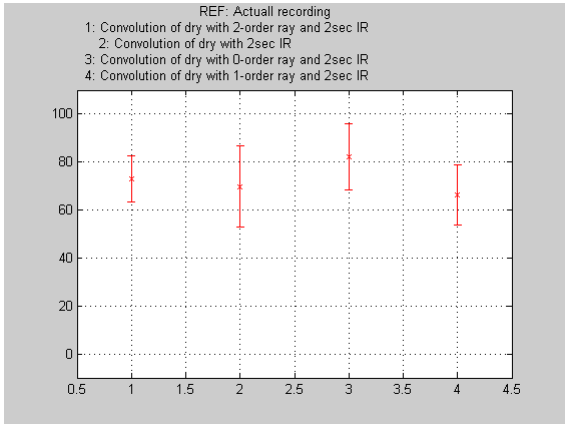


Figure 5.7: Effect of Orders of Reflections

source location. Using the direct arrival and the first order reflections offers a closer recreation. It also shows that we can use the reverberation tails measured at B to recreate the late reverb for A.

The figure on the right in Fig. 5.9 suggests that listeners are rather tolerant regarding the inaccuracy in the absorption filter estimation, comparing with the significant impact by inaccurate reverberation tails in Fig. 5.8. However, it does show that the early reflection renderer generated from an vastly different room degrades the overall quality to a certain degree.

Preference on Order of Reflections The preference evaluation results shown in Fig. 5.10 on this topic do not provide conclusive results, meaning that the listener pre-

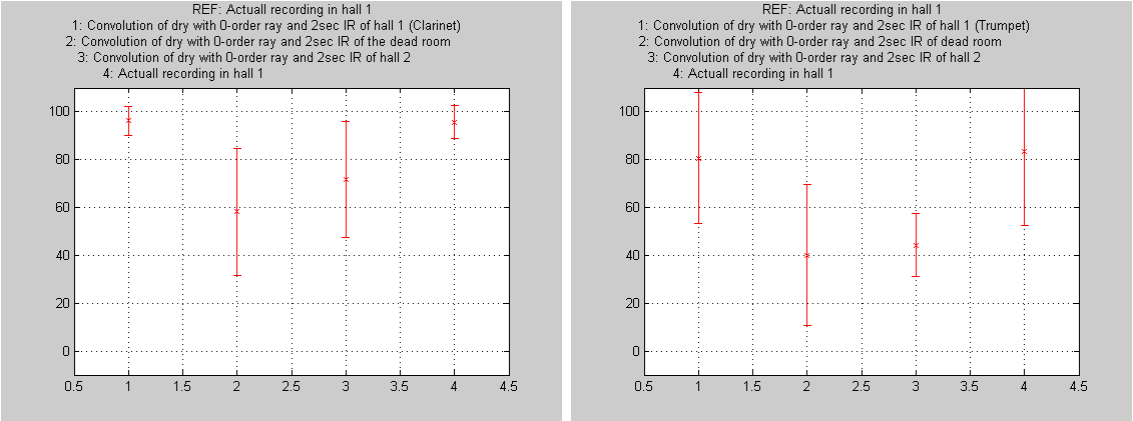


Figure 5.8: Mismatching Late Reverberation

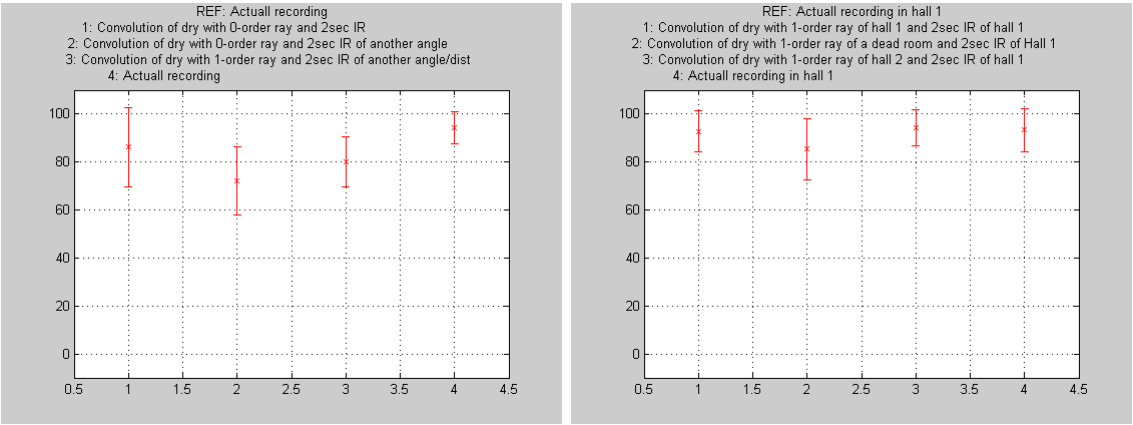


Figure 5.9: Mismatching Early Reflections

fer equally amongst pure convolution, convolution with direct sound and convolution with direct sound and first order reflections.

5.2.4 Summary

In summary, the evaluation results clearly demonstrated that the proposed system is superior in retaining the spatial impression and creating a enjoyable listening experience than artificial reverberators.

The above results and analysis also justify the fundamental assumptions of the proposed hybrid system, i.e.

1. late reverberation determines the overall acoustic impression and has little effect

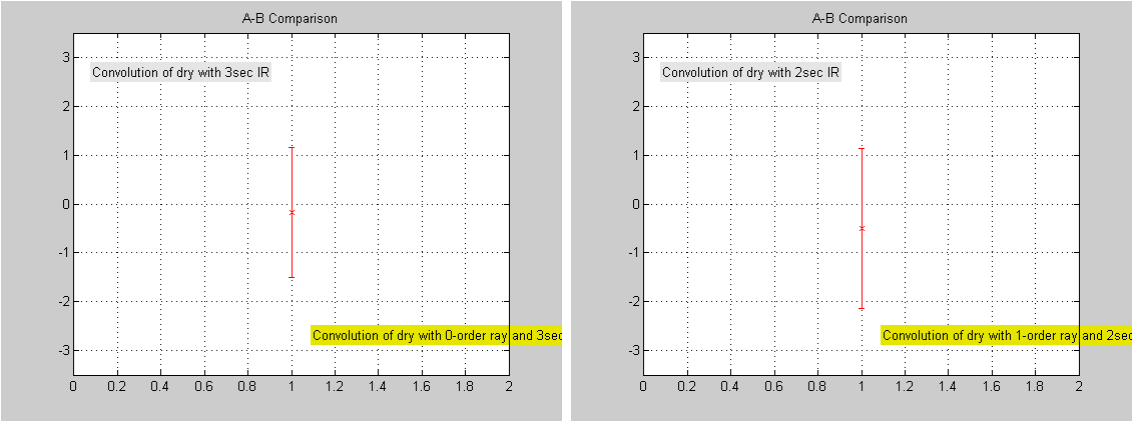


Figure 5.10: Preference on Order of Reflections

on the localization;

2. the MMRIR acquired at one "good" position in the target space can be used to simulate the late reverb at other positions in the same space;

3. early reflections affect greatly the source localization and has limited effect on overall acoustic impression.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, the principles and requirements of spatial sound rendering, as well as the existing solutions were reviewed. Then techniques for acquiring multichannel room impulse responses were discussed. Built upon the MMRIRs, a new spatial sound rendering system has been developed. The new system uses a hybrid model that models only the direct sound and early reflections individually using the image-source method and synthesizes the late reverberation using a set of filters derived from the MMRIR. The image-source model is built upon the parameters estimated from MMRIR. Randomization can be applied to these parameters to simulate diffraction. The multichannel reverberation tails are created by filtering the input signal with (optionally) decorrelated MMRIR tails. Compared with existing solutions, the proposed system offers the following key benefits:

1. the model is built upon RIR measurement which is a true reflection of physical acoustics in the measured room;

2. it is can be easily extended to produce a new spatial impression - only multichannel RIR measurements are needed;

3. it is scalable and flexible in that its quality and complexity can be controlled easily;

4. it is able to simulate arbitrary source-listener configurations.

5. the current system focuses on the 5-channel multichannel reproduction, but it may be extended to any speaker or headphone configuration, with or without the corresponding MMRIR.

The proposed system was implemented on PC to demonstrate its capability of performing real-time rendering. Formal subjective evaluation has been carried out and the results fully demonstrated the effectiveness and potential of this hybrid approach.

6.2 Future Work

There are several areas where our system may be improved. The MMRIR acquisition system may be improved to reduce the SNR and distortion. Post-processing can be applied to the measurements to isolate noise and nonlinear distortion [83]. Real-time analysis and visualization tools are very helpful in finding defects in the measurements and allowing for instant adjustments. In the analysis of the MMRIR, it is possible to refine the estimation of the absorption filters, for example, to estimate an individual surface absorption filter for each wall using the first few reflections in the RIR. The analysis process could be further streamlined if the direct sound and early reflections in the MMRIRs can be identified and removed automatically.

In the early reflection renderer, it may be necessary to expand the image source model to handle complex room geometries. Efficient ray tracing or image source algorithms for computer graphics have been available for many years and can be adapted for our system. It would be interesting to study the differences among the reverb tails in a set of MMRIRs, and see if a proper decorrelator can be designed based on the findings. Because the tail of RIR is normally very long, fast convolution using FFT may still exceed the available computing capacity in some cases such as rendering multiple spaces simultaneously. More efficient methods, e.g. IIR approximation [114] or the Common-Acoustical-Pole Zero model [115], are being investigated. The current implementation only renders one sound source and needs to be revised to handle multiple sources. It would be very interesting to see how this system can be adapted so that it efficiently runs on multi-core processors or GPUs.

The subjective evaluation may be expanded to incorporate more materials, to use different MMRIRs and to expand the participant group. It would be very interesting conduct the evaluation in various listening environments and speaker configurations, for example, a music production control room with stereo speakers and a cinema the-

ater with a large number of surround speakers. These would reveal how the acoustic properties of the listening environment affect our rendering quality. Another potential topic of the subjective evaluation is to compare the rendering quality using the decorrelated mono reverb tail with the quality using the multi-channel reverb tails.

The proposed system lays out a framework for high quality spatial rendering. It is not, yet, a complete 3d audio renderer that can be used in interactive applications such as gaming, as it is missing a few key elements such as sound source directivity, occlusion and obstruct, diffraction and etc. However, these elements can be added within the framework to form a complete 3d audio system.

Appendix A

User Study Procedures and Results

A.1 Test Plan

This is **NOT** to be disclosed to the participants. During the tests, trials and conditions will be randomized.

A.1.1 Training

1. Real Recording vs. Dry:
   Purpose: To get the users familiarized with the interface and the range of possible quality differences.
   Conditions:

   - Sur Rec.
   - Sur Rec. as 100 (highest possible score)
   - Dry as 0 (lowest possible score)

A.1.2 Similarity Trials (MUSHRA)

In the following, "Signals" will be referred as "Conditions". Each "condition" will have a length of 20 second, as recommended by ITU-R Recommendation BS.1534-1. The ARL STEP program takes up to four test conditions plus the reference. In our case, all similarity trials have 4 conditions (5 including the reference). If a trial is to evaluate less than 4 conditions, a hidden reference is inserted.

1. RayIR vs. Artificial Reverb:
   Purpose: To determine if the Ray+IR and IR is BETTER than the artificial reverberators.
   Conditions
   - REFERENCE Actual recording
   - Audition 1.5 QuickVerb (Artificial)
   - Audition 2.0 Studio Reverb (Artificial)
   - Convolution of dry with 1-order ray and 3sec IR
   - Convolution of dry with 3sec IR
2. RayIR vs. Convolutive Reverb:
   Purpose: To determine how does the Ray+IR and IR compare to the convolutive reverberators.
   Conditions
   - REFERENCE Actual recording
   - Audition 2.0 Full Reverb (Convolutive)
   - Audition 2.0 Reverb (Convolutive)
   - Convolution of dry with 3sec IR
   - Convolution of dry with 0-order ray and 3sec IR
3. RayIR vs. Ray:
   Purpose: To determine how does the Ray+IR compare to Ray only.
   Conditions
   - REFERENCE Actual recording
   - Convolution of dry with 2-order ray and 2sec IR
   - Convolution of dry with 2sec IR
   - Convolution of dry with 0-order ray and 2sec IR
   - Convolution of dry with 1-order ray and 2sec IR
4. Full Length Tail vs. Truncated Tail (hall):
   Purpose: To determine the length of IR tail is needed to provide full perceptual quality.
   Conditions

- REFERENCE Convolution of dry with 3sec IR
- Convolution of dry with 0.4sec IR
- Convolution of dry with 0.8sec IR
- Convolution of dry with 1.2sec IR
- Convolution of dry with 1.6sec IR

5. Full Length Tail vs. Truncated Tail (impulsive sound in a dead room):
   Purpose: To determine the length of IR tail is needed to provide full perceptual quality.
   Conditions

   - REFERENCE Convolution of dry with 2sec IR
   - Convolution of dry with 0.4sec IR
   - Convolution of dry with 0.6sec IR
   - Convolution of dry with 0.8sec IR
   - Convolution of dry with 1.0sec IR

6. (RayIR) Hall A vs. Hall B (Clarinet):
   Purpose: To determine if the listener is able to differentiate hall A from B.
   Conditions

   - REFERENCE Actual recording in hall 1
   - Convolution of dry with 0-order ray and 2sec IR of hall 1
   - Convolution of dry with 0-order ray and 2sec IR of the dead room
   - Convolution of dry with 0-order ray and 2sec IR of hall 2
   - Actual recording in hall 1

7. (RayIR) Hall A vs. Hall B (Trumpet):
   Purpose: To determine if the listener is able to differentiate hall A from B.
   Conditions

   - REFERENCE Actual recording in hall 1
   - Convolution of dry with 0-order ray and 2sec IR of hall 1
   - Convolution of dry with 0-order ray and 2sec IR of the dead room

- Convolution of dry with 0-order ray and 2sec IR of hall 2
- Actual recording in hall 1

8. Static tail vs. Real tail:
   Purpose: To determine if using static IR tail is okay.
   Conditions - Hall

   - REFERENCE Actual recording in hall 1
   - Convolution of dry with 0-order ray and 2sec IR of hall 1
   - Convolution of dry with 0-order ray and 2sec IR of the dead room
   - Convolution of dry with 0-order ray and 2sec IR of hall 2
   - Actual recording in hall 1

   Conditions - impulsive sound in a dead room

   - REFERENCE Actual recording in hall 1
   - Convolution of dry with 0-order ray and 2sec IR of hall 1
   - Convolution of dry with 0-order ray and 2sec IR of the dead room
   - Convolution of dry with 0-order ray and 2sec IR of hall 2
   - Actual recording in hall 1

   Note: Repeating this on position B,C,D, is to see what will happen if we use a static IR for "moving" sources (B,C,D). This three trials are to be run consecutively.

9. Mismatch Ray with IR:
   Purpose: To determine if the hall impression is affected by early reflections.
   Conditions

   - REFERENCE Actual recording in hall 1
   - Convolution of dry with 1-order ray of hall 1 and 2sec IR of hall 1
   - Convolution of dry with 1-order ray of a dead room and 2sec IR of Hall 1
   - Convolution of dry with 1-order ray of hall 2 and 2sec IR of hall 1
   - Actual recording in hall 1

A.1.3 Preference trials (A-B Comparison)

NOTE: There trials are to evaluate which one is better between A and B.

1. RayIR vs. Convolutive Reverb:
   Purpose: To determine how does the Ray+IR compare to the convolutive reverberators.
   Conditions
   - Audition 2.0 Reverb (Convolutive)
   - Convolution of dry with 3sec IR
2. RayIR vs. Artificial Reverb:
   Purpose: To determine if the Ray+IR is BETTER than the artificial reverberators.
   Conditions
   - Audition 1.5 QuickVerb (Artificial)
   - Convolution of dry with 1-order ray and 3sec IR
3. RayIR vs. Ray:
   Purpose: To determine how does the Ray+IR compare to Ray only.
   Conditions
   - Convolution of dry with 3sec IR
   - Convolution of dry with 0-order ray and 3sec IR
4. RayIR vs. Ray:
   Purpose: To determine how does the Ray+IR compare to Ray only.
   Conditions
   - Convolution of dry with 2sec IR
   - Convolution of dry with 1-order ray and 2sec IR
5. Static tail vs. Real tail:
   Purpose: To determine if using static IR tail is okay.
   Conditions
   - Convolution of dry with 0-order ray and 2sec IR of another angle

- Convolution of dry with 1-order ray and 2sec IR of another angle/dist

6. Static tail vs. Real tail #1:
   Purpose: To determine if using static IR tail is okay.
   Conditions

   - Convolution of dry with 2sec IR
   - Convolution of dry with 1-order ray and 2sec IR of another angle/dist

A.2 Instructions for Participants

Listeners **must** read this prior to their participation in the test!

A.2.1 Introduction

We are interested in assessing the performance of a Spatial Sound Rendering System that permit recreating the acoustic impression of certain venue on a multichannel speaker system. All test items that you will hear are 5.1 channel "surround-sound" signals, with 3 front channels Left, Center, Right (L, C, R), and 2 surround channels Left-surround, Right-surround (LS and RS) (towards the rear) and a subwoofer channel.

This research will be using standardized software instrument - **Subjective Training and Evaluation Program (STEP)** by Audio Research Labs.

A.2.2 Training phase

You will be presented three multichannel audio signals, including the reference signal, one original recording (highest quality) and one close mic (dry) signal (lowest quality). In this training segment you will be able to switch back and forth at will. Please listen to the training signals to learn how the reference sounds, and then evaluate the various other signals appropriately, identifying and providing them a grade as explained in the test instructions. In the course of listening to these signals, please pay particular attention to:

1. The location of particular sounds
2. The presence of reverberation

Figure A.1: Training Session

3. The overall quality including, but not limited to, tonal quality, spatial quality and clarity.

Bear in mind that any change from the original is to be considered an impairment.

A.2.3 Testing phase 1 - Similarity tests

In the first session (see Fig. A.2), you will be listening to a reference sound signal (REF) and a number of other sound signals (A-D). You are to determine the **similarity** of A-D to REF and mark the **similarity** on the scale of 0 to 100. Upon finish, click NEXT to continue.

You are asked to judge the **Basic Audio Quality** of the versions of the test items in each trial. This attribute is related to **any and all differences** between the reference and the tests item, including **soundstage rendition**, **envelopment**, **reverberation**, **distortion**, and **bandlimiting**, such that **any difference** between the reference and the test items is to be **considered an impairment**. The assessment is to be done on a scale from **0 to 100**.

There might be trials where REF plays different musical content from cond. ABCD. In these cases, please focus on the **impression of acoustic space**.

A.2.4 Testing phase 2 - Perference tests

In the second session (see Fig. A.3), you will be listening to two sound signals (A,B). You are to determine which one you **prefer** and mark the **preference** on the scale

of -3 to 3. Upon finish, click NEXT to end to test.

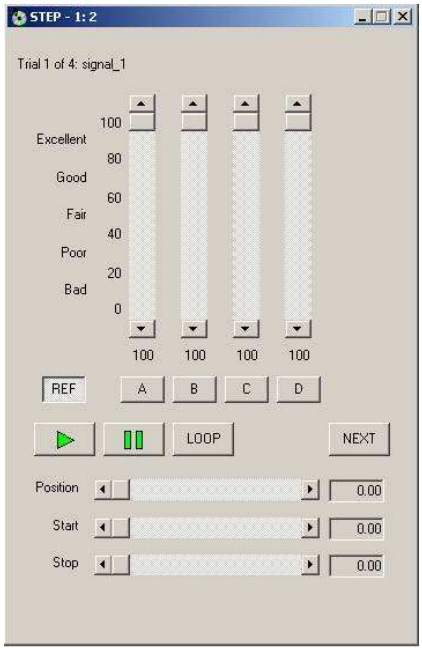


Figure A.2: Mushra Session

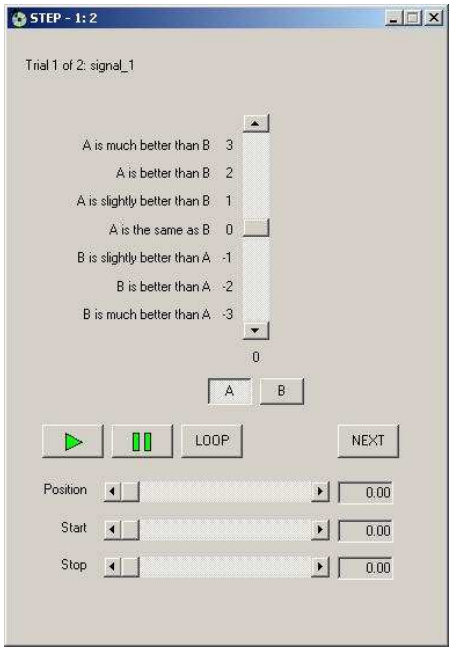
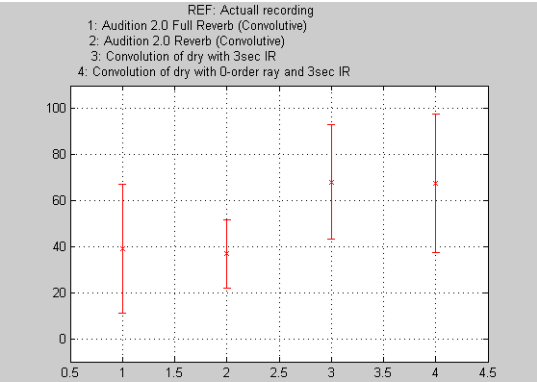
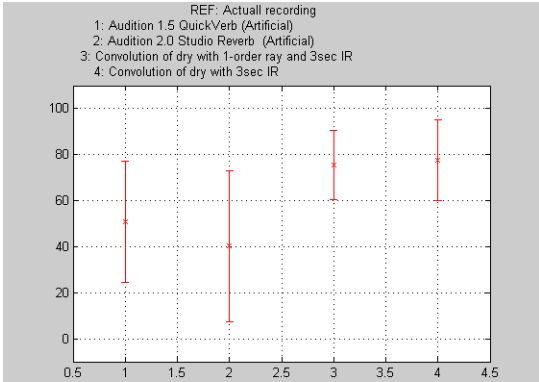
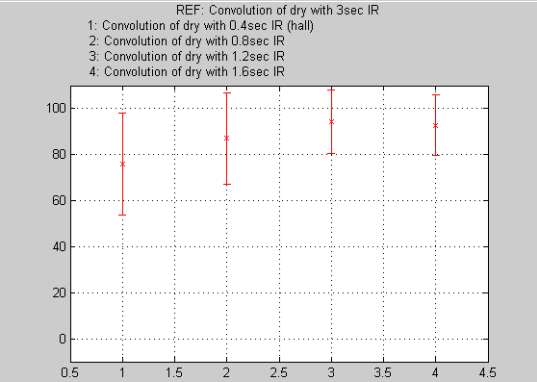


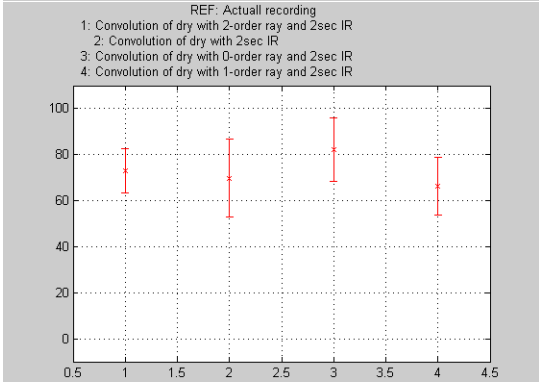
Figure A.3: A-B Session

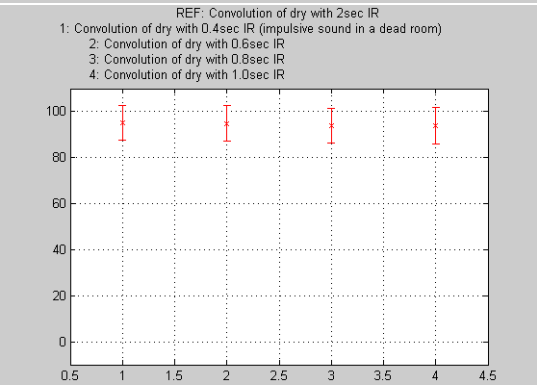
A.3 Complete Evaluation Results

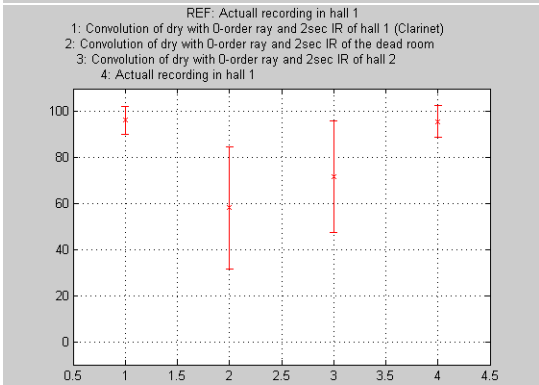

REF: Actuall recording
1: Audition 2.0 Full Reverb (Convolutive)
2: Audition 2.0 Reverb (Convolutive)
3: Convolution of dry with 3sec IR
4: Convolution of dry with 0-order ray and 3sec IR

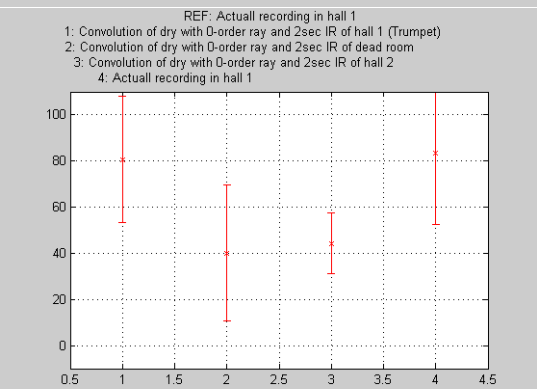

REF: Actuall recording
1: Audition 1.5 QuickVerb (Artificial)
2: Audition 2.0 Studio Reverb (Artificial)
3: Convolution of dry with 1-order ray and 3sec IR
4: Convolution of dry with 3sec IR

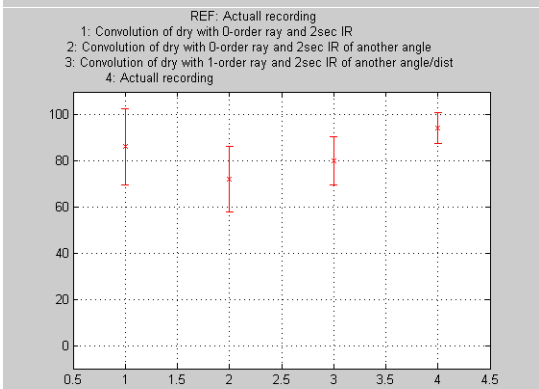

REF: Convolution of dry with 3sec IR
1: Convolution of dry with 0.4sec IR (hall)
2: Convolution of dry with 0.8sec IR
3: Convolution of dry with 1.2sec IR
4: Convolution of dry with 1.6sec IR

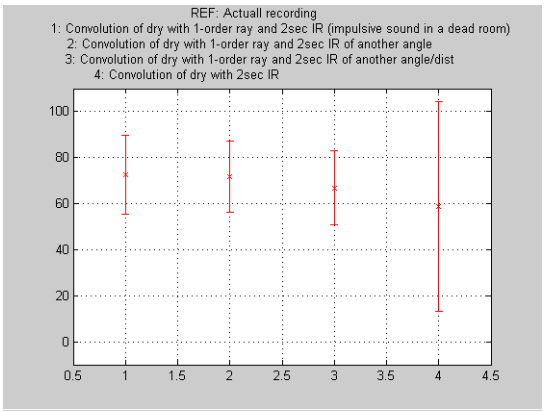

REF: Actuall recording
1: Convolution of dry with 2-order ray and 2sec IR
2: Convolution of dry with 2sec IR
3: Convolution of dry with 0-order ray and 2sec IR
4: Convolution of dry with 1-order ray and 2sec IR

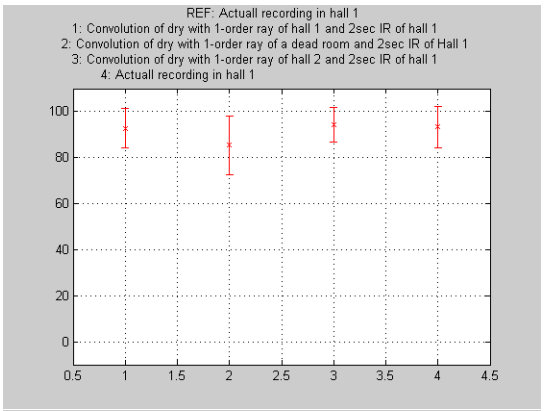

REF: Convolution of dry with 2sec IR
1: Convolution of dry with 0.4sec IR (impulsive sound in a dead room)
2: Convolution of dry with 0.6sec IR
3: Convolution of dry with 0.8sec IR
4: Convolution of dry with 1.0sec IR

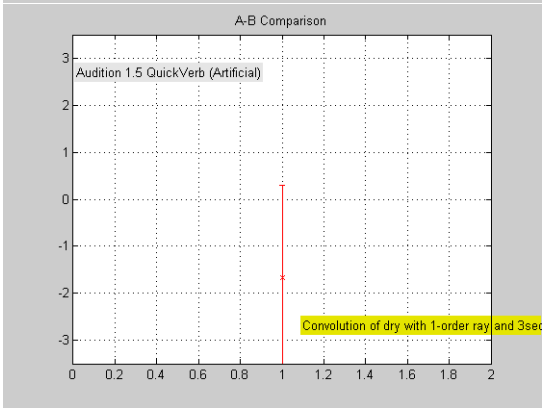

REF: Actuall recording in hall 1
1: Convolution of dry with 0-order ray and 2sec IR of hall 1 (Clarinet)
2: Convolution of dry with 0-order ray and 2sec IR of the dead room
3: Convolution of dry with 0-order ray and 2sec IR of hall 2
4: Actuall recording in hall 1

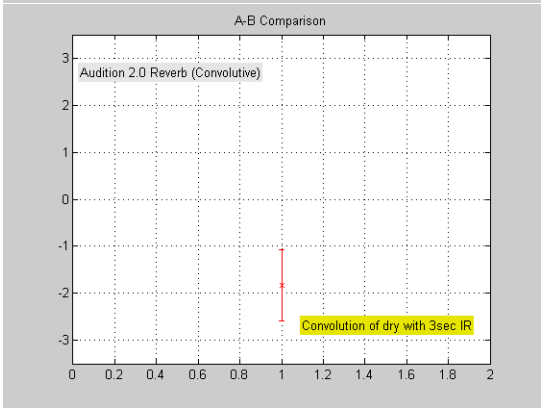

REF: Actuall recording in hall 1
1: Convolution of dry with 0-order ray and 2sec IR of hall 1 (Trumpet)
2: Convolution of dry with 0-order ray and 2sec IR of dead room
3: Convolution of dry with 0-order ray and 2sec IR of hall 2
4: Actuall recording in hall 1

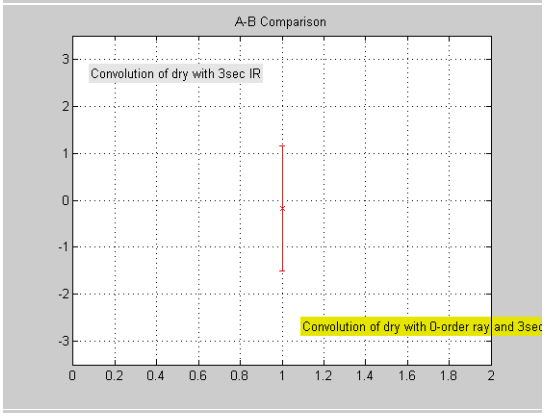

REF: Actuall recording
1: Convolution of dry with 0-order ray and 2sec IR
2: Convolution of dry with 0-order ray and 2sec IR of another angle
3: Convolution of dry with 1-order ray and 2sec IR of another angle/dist
4: Actuall recording

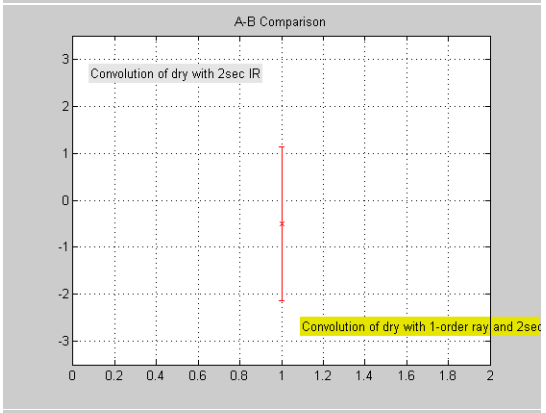

REF: Actuall recording
1: Convolution of dry with 1-order ray and 2sec IR (impulsive sound in a dead room)
2: Convolution of dry with 1-order ray and 2sec IR of another angle
3: Convolution of dry with 1-order ray and 2sec IR of another angle/dist
4: Convolution of dry with 2sec IR

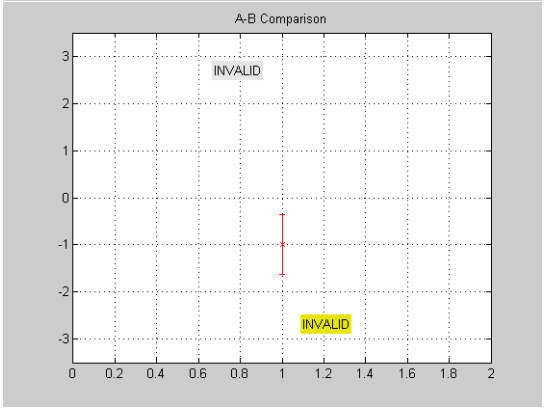

REF: Actuall recording in hall 1
1: Convolution of dry with 1-order ray of hall 1 and 2sec IR of hall 1
2: Convolution of dry with 1-order ray of a dead room and 2sec IR of Hall 1
3: Convolution of dry with 1-order ray of hall 2 and 2sec IR of hall 1
4: Actuall recording in hall 1

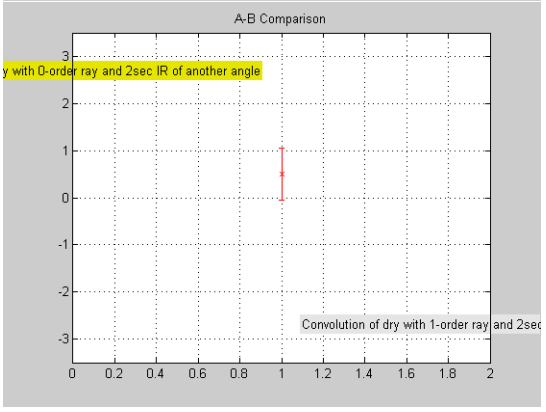

A-B Comparison
Audition 1.5 QuickVerb (Artificial)
Convolution of dry with 1-order ray and 3sec


A-B Comparison
Audition 2.0 Reverb (Convolutive)
Convolution of dry with 3sec IR


A-B Comparison
Convolution of dry with 3sec IR
Convolution of dry with 0-order ray and 3sec


A-B Comparison
Convolution of dry with 2sec IR
Convolution of dry with 1-order ray and 2sec


A-B Comparison
INVALID
INVALID


A-B Comparison
y with 0-order ray and 2sec IR of another angle
Convolution of dry with 1-order ray and 2sec

Appendix B

Software Package

For the readers of this thesis to be able to experience the proposed spatial sound rendering solution, a complete software package has been development using Matlab and C++. This is an integral part of this thesis because one of the goals is to build a real-time system.

The functionality of components of this software package and the usage of each component are given in the following sections, followed by a short tutorial demonstrating the steps need to be followed.

Complete source codes are available by request.

This software package consists of three main modules, as listed below.

- (Matlab) `banff200610/config.m` calculates and generates IR and PRESET from recorded wave files.
- (Win32 console) `ra3dcon44/48/96.exe` is a command line application that generates a multichannel wave file (using `WAVE_EXTENSIBLE` format) from a mono wave file based on the settings in a configuration file.
- (Win32 GUI) `ra3dui44/48/96.exe` is the real-time demo program that takes a mono wave file or a mono audio input and renders it into 5.1 audio.

B.1 Matlab

The Matlab scripts calculate the room impulse responses from the recorded sweep signals and generate PRESETs from these impulse responses. The scripts require the user to specify a number of options in order to work properly. The scripts also

require the input wave files to be named in a descriptive way so that the files can be recognized correctly.

B.1.1 User Specified Options

Setting Up Directories

```
% the root dir of the matlab scripts.
dir_research = 'D:\\yli\\RayAudio\\subvn\\ra3d\\research\\';

% where the input wave files are
dir_rec = 'F:\hall\';

% where to save the assemblied surround recording
% in the format of extensible wave files
dir_sav = 'D:\yli\RayAudio\recording\';

% where to save the generated impulse responses and PRESETs
dir_dat = 'D:\yli\RayAudio\recording\';
```

Setting Up Session Information

These options are specific for each recording session, such as room dimensions, microphone array position and configuration, and so on.

```
% the signature string in the wave file names of this session
pre_name = 'uvic102306_msmts'; % for hall

% [USER DEFINED] signature string to be used in all generated files.
pre_irname = 'banff20061023studio'; % for hall

% room dimension of the recording venue
room_dim = [5.0, 5.0, 3.0];

% position of the center of the microphone array
mic_pos = [0.0, 0.0, -0.4];

% mic pattern A and B
mic_pat = [5.0, 7.0]; %supercard

% radius of microphone array
mic_r = 0.15;

% source height
src_z = -1.5;

%channel prefix string of the wave files
```

```
channels = strvcat('Sweep 1', 'LEFT', 'RIGHT', 'CENTRE', 'LS', 'RS', 'SPOT');

% number of sweep cycles
n_totalcycle = 2;

% length of each sweep cycle
l_sweep = 260000;

% [USER DEFINED] length of match filter to use
l_mf = 1100;

% starting and ending point of audio segments in the input wave files
psamp =[1 44100*14;
        44100*18 44100*32;
        44100*36.5 44100*66.5; %44100*76.5;
        44100*83.8 44100*113.8; %44100*143.8;
        44100*150 44100*180;
        44100*183.8 44100*213.8;%44100*254.8;
        ];

% [USER DEFINED] name of audio segments in the input wave files
name_seg = ['lin';'log';'dru';'trp';'din';'cla'];

% [USER DEFINED] name of channels
name_chn = ['sw';'fl';'fr';'fc';'sl';'sr';'sp'];
```

Setting Up Parameters

These are the parameters used to calculate impulse responses and PRESETs and, in turn, controls the audio rendering.

```
% the target sampling rate
ir_fs = 44100;

% length of impulse responses
n_ir = 2*44100;

% threshold for removing pikes in recorded sweep
thre_pikes = 1.2;

% the original signal level, can be approximated by
% un-attenuated close mic signal level.
closemic_level = 0.5;

% reverb time level in dB, e.g -60 for RT60
rt_level = -70;

% IIR filter order for approximating absorption filter
iir_order = 2;
```

B.1.2 Input File Naming Convention

The wave files need to be named as follows

`[channel prefix]_[signature]_[position]_*.wav,`

for example,

```
Sweep 1_uvic102306—msmts 0 117 (7)_##008##__{816D1D62—CD71—4F97—B831—1AE4CFEABCD9}.wav}.
```

B.1.3 Content of a Preset

The results of impulse response analysis are saved in a matlab data file what contains the the following fields. This file can be loaded within Matlab or by the C++ library `RA3DUtil`.

Name	Type	Notes
`iir_order`	int	order of absorption filters
`filtlen`	int	length of impulse responses
`samp_rate`	float	sampling rate
`ref_level`	float	reference level $\sum \lvert firstpeaks \rvert$
`ref_dist`	float	reference distance where attenuation is 1
`initial_delay`	int	theoretical delay preceding the first peak
`src_pos`	vector-3	actually source position of ir measurement
`mic_pat`	vector-2	micphone pattern $a + b \sin\theta$
`mic_pos`	vector-3	actually mic position of ir measurement
`room_dim`	vector-3	actually room dimensions

`hf_fl`
`hf_fr`
`hf_fc`
`hf_sl`
`hf_sr`
	vector	contains IR of length `filtlen`

`ab`
`aa`
`wb`
`wa`
	vector	absorption filter coefficients of length `iir_order+1`

`preset_name`	string	name of the preset

Table B.1: Content of a PRESET

B.2 Windows Applications

Two Windows applications were developed to demonstrate the real-time capability. C++ programming language was used instead of Matlab because of Matlab's low efficiency and incapability of real-time audio streaming. With proper hardware configuration, the applications are able to process up to 1.2 seconds of 5-channel reverb tail.

B.2.1 Console Program: `RA3DCon44/48/96.exe`

This is the command line program. It requires one argument which is the name of the configuration file. For example, `ra3dcon config.cfg`. The configuration file looks like the following,

```
#input
file_in = "drytrumpet.wav"
file_preset = "preset_banff20061022hall_a324d247_lin_2_87181.mat"

#output
file_6c = "drytrumpet6.wav"

#parameters
lis_n_damping = 100
lis_n_reverbmix = 100
lis_n_filterlen = 0
lis_n_downfactor = 0
lis_n_mode = 0
lis_v_room = 7.0, 10.0, 3.0
lis_v_pos = 0.0, 0.0, -0.4
lis_v_ori = 0.0, 0.0, 0.0

#parameters
src_n_rayorder = 0
src_v_pos = -3.7, 5.1, -1.5

#parameters
misc_framelimit = 500
```

Some of the fields are not actually in use currently, including `lis_n_downfactor`, `lis_n_reverbmix` and `lis_n_damping`. The in-use fields are listed in Table. B.2.

Name	Type	Notes
`lis_n_filterlen`	int	percentage of the ir length to use
`lis_n_mode`	int	1: no reflections 0: normal
`misc_framelimit`	int	no.s of audio frame to be processed (size of the frame/block is defined in `ra3d.h`

Table B.2: Fields in the configuration file

B.2.2 Demo(GUI) Program

The demo program renders one mono wave file or one mono input, which could be whatever the recording source of the PC audio card takes, to 5.1 surround sound. To use the duplex mode to render mono 'live' input, proper recording source must be set by checking proper channel in the recording volume panel, as shown in Fig. B.1 The

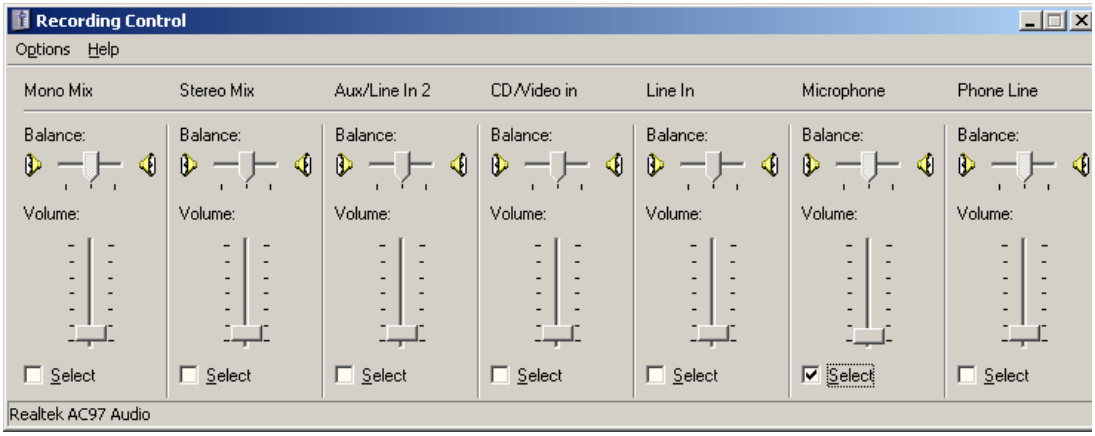


Figure B.1: Recording Source Selection

system requirements to run this program are listed in Table. B.3.

- AMD Athlon 64 X2 3500+ or Intel equivalent;
- 512M Ram;
- Multichannel audio card with **DirectSound** support.

Table B.3: System Requirements

The main window of the demo application is shown in Fig. B.2. The head icon at the center of this window represents the listener, while the helicopter icon represents

the sound source. The sound source can be moved by clicking the left mouse button inside this window. User can also hold the `CTRL` key when clicking to move the source source slowly to the target position. This window can be re-sized and the dimension of the room will change accordingly. However, resizing is NOT recommended because it will not be consistent with the room dimensions in the PRESET.

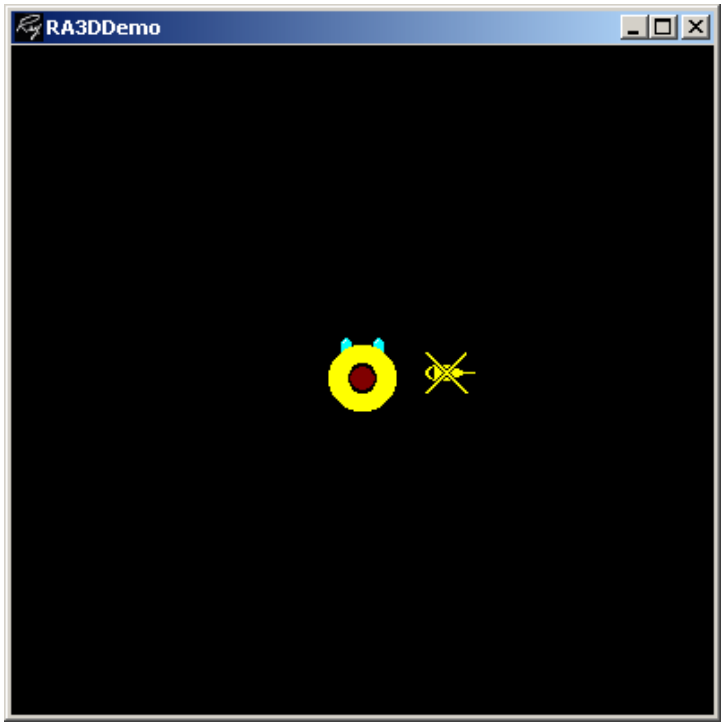


Figure B.2: Main Window

The control dialog is shown in Fig. B.3. This dialog is responsible for selecting inputs, loading the input files, adjusting parameters as well as displaying activities inside the core module.

The details of each control is listed in Table. B.4.

Recommended steps to use this program are,

1. To load a preset - **9**;
2. To select an input - **17-19**;
3. To adjust parameters.

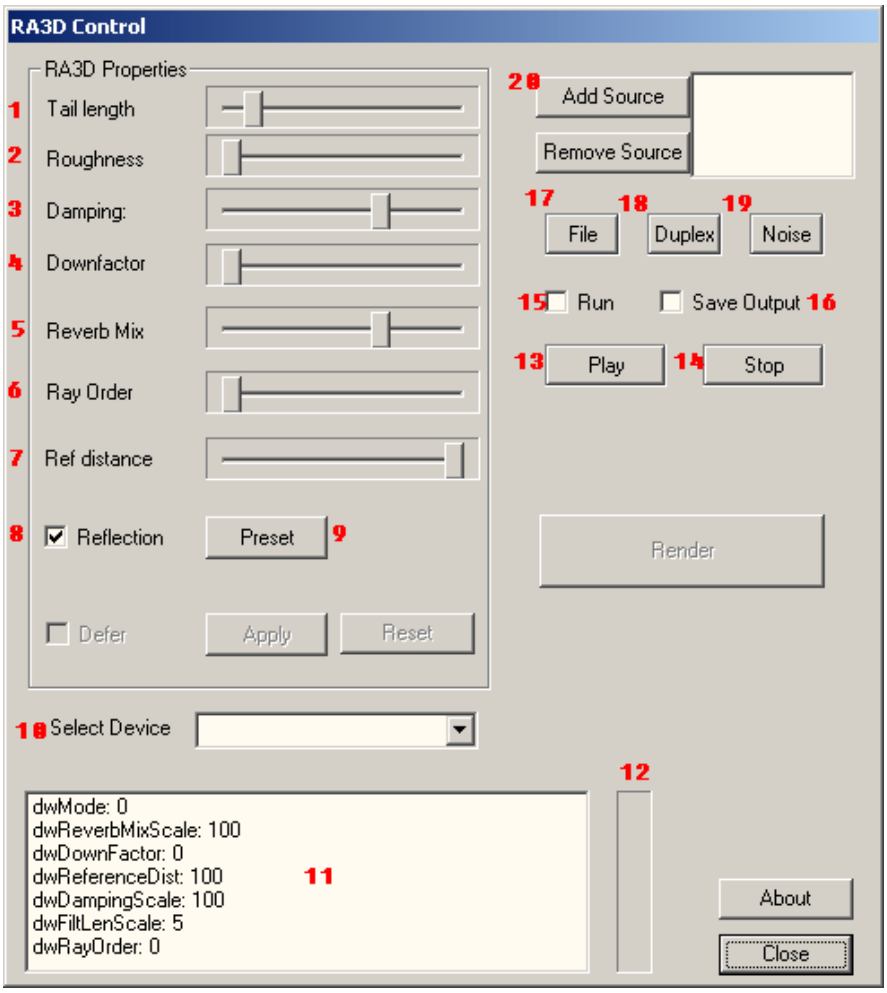


Figure B.3: Control Dialog

B.2.3 C++ Interface and Libraries

The core functionality of AR3D are implemented in a self-contained C++ library. For detailed information, the reader is referred to `ra3d.h` and `cra3dutil.h`.

No.	Name	Notes
1	Tail Length	Percentage of the IR tail to use
2	Roughness	Randomization factor
6	Ray Order	
7	Reference Distant	
8	Reflections	Whether to use reflections
9	Preset	To load preset
11	Message	
12	CPU Usage	
13,14	Play, Stop	
15	Run	When checked, the source will circle around the listener
16	Save Output	
17,18,19	Select Inputs	"noise" is the pink noise channel test
Rest	-	No in use

Table B.4: AR3D Demo Control

Bibliography

[1] M. Kleiner, B.-I. Dalenback, and P. Svensson, "Auralization - an overview," *J. Audio Eng. Soc*, vol. 41, no. 11, pp. 861–875, 1993.

[2] T. Lokki, *Physically-based Auralization: Design, Implementation, and Evaluation*. PhD thesis, HUT, 2002.

[3] K. H. Kuttruff, "Auralization of impulse responses modeled on the basis of ray-tracing results," *J. Audio Eng. Soc*, vol. 41, no. 11, p. 876, 1993.

[4] T. D. Rossing, *The Science of Sound*. Addison-Wesley, Reading, UK, 1990.

[5] W. G. Gardner, *3-D Audio Using Loudspeakers*. PhD thesis, MIT Media Lab, 1997.

[6] J. J. Smaldino, C. C. Crandell, B. M. Kreisman, A. B. John, and N. V. Kreisman, *Audiology Treatment*, ch. Room Acoustics for Listeners with Normal Hearing and Hearing Impairment. Thieme Medical Publishers, 2007.

[7] L. S. Lloyd, *Music and Sound*. Ayer Publishing, 1970.

[8] W. G. Gardner, *Applications Of Digital Signal Processing To Audio And Acoustics*, ch. Reverberation Algorithms, p. 85. Kluwer Academic Publishers, 2002.

[9] S. Lehman, "Reverberation." harmony-central.com. `http://www.harmony-central.com/Effects/Articles/Reverb/`.

[10] H. J. Fredrics, "Bouncing off walls," *Electronic Musician*, 2009.

[11] R. O. Duda, "3-D audio for HCI." `http://interface.cipic.ucdavis.edu/CIL_tutorial/3D_home.htm`, 2000.

[12] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," *J. Audio Eng. Soc*, vol. 49, pp. 231–249, 2001.

[13] J. Speigle, J.M.; Loomis, "Auditory distance perception by translating observers," in *Proceedings of the IEEE Symposium on Research Frontiers in Virtual Reality*, 1993.

[14] C. Sheeline, *An Investigation of the Effects of Direct and Reverberant Signal Interactions on Auditory Distance Perception*. PhD thesis, Stanford University, 1982.

[15] C. J. Plack, *The Sense of Hearing*. Psychology Press, 2005.

[16] S. Devore, A. Ihlefeld, K. Hancock, B. Shinn-Cunningham, and B. Delgutte, "Accurate sound localization in reverberant environments is mediated by robust encoding of spatial cues in the auditory midbrain," *Neuron*, vol. 62, pp. 123–134, 2009.

[17] G. Theile, "Natural 5.1 music recording based on psychoacoustic principals," in *Proc. AES 19th International Conference*, 2001.

[18] D. Rutter, "Head to head - headphones versus speakers," *Atomic: Maximum Power Computing.*, 2008.

[19] W. G. Gardner, "3D audio and acoustic environment modeling," *Wave Arts Inc. Whitepaper*, 1999. `http://www.wavearts.com`.

[20] H. L. Han, "Measuring a dummy head in search of pinna cues," *J. Audio Eng. Soc*, vol. 42, pp. 15–37, 1994.

[21] F. Filipanits Jr., "Design and implementation of an auralization system with a spectrum-based temporal processing optimization," Master's thesis, University of Miami, 1994.

[22] J. Pikover, "5 surround sound headsets." Tom's Guide US, 2009. `http://www.tomsguide.com/us/Surround-Sound-Headsets,review-1357.html`.

[23] "Which computer speaker set do you have?." online poll, 2006.

[24] "ITU-R BS.775-1: Multichannel Stereophonic Sound System with and without Accompanying Picture."

[25] C. Tonnesen and J. Steinmetz, "3D sound synthesis, encyclopedia of virtual environments." `http://www.hitl.washington.edu/scivw/EVE/I.B.1.3DSoundSynthesis.html`, 1993.

[26] H. Kuttruff, "Sound field prediction in rooms," in *Proc. 15th Int. Congr. Acoust.*, 1995.

[27] D. A. Burgess, "Techniques for low cost spatial audio," in *ACM Symposium on User Interface Software and Technology*, pp. 53–59, 1992.

[28] J. Jot, "Efficient models for reverberation and distance rendering in computer music and virtual audio reality," in *Proc. 1997 Int. Computer Music Conf.*, 1997.

[29] Aureal Corporation, "3-D audio primer." HeadWize Technical Library, 1998. `http://www.headwize.com/tech/aureal1_tech.htm`.

[30] M. R. Schroeder, "Natural-sounding artificial reverberation," *J. Audio Eng. Soc.*, vol. 10, no. 3, 1962.

[31] L. Savioja, J. Huopaniemi, T. Lokki, and R. Vaananen, "Virtual environment simulation - advances in the DIVA project," in *Proc. Int. Conf. Auditory Display*, 1997.

[32] N. Tsingos, E. Gallo, and G. Drettakis, "Perceptual audio rendering of complex virtual environments," *ACM Transactions on Graphics*, vol. 23, pp. 249–258, 2004.

[33] S. Heise, M. Hlatky, and H. Bremen, "Automatic adjustment of off-the-shelf reverberation effects," in *AES 126th Convention*, 2009.

[34] "ISO/IEC JTC1/SC29/WG11: Information technology-coding of audio-visual objects. part1: Systems."

[35] A. Simeonov, G. Zoia, , and R.-L. Garcia, "Rendering of advanced 3D room models by enhanced application programming interfaces," in *AES 114th Convention*, 2003.

[36] V. Pulkki, *Spatial Sound Generation and Perception by Amplitude Panning Techniques*. PhD thesis, Helsinki University of Technology, 2001.

[37] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, 2001.

[38] J. Moorer, "About this reverberation business," *Computer Music Journal*, vol. 3, pp. 13–28, 1979.

[39] M. R. Schroeder, "New method of measuring reverberation time," *J. Audio Eng. Soc.*, 1965.

[40] E. Doering, "Reverberation." `http://cnx.org/content/m15471/latest/`.

[41] Z. Rafii and B. Pardo, "Learning to control a reverberator using subjective perceptual descriptors," in *10th International Society for Music Information Retrieval (ISMIR 2009)*, 2009.

[42] J.-M. Jot and A. Chaigne, "Digital delay networks for designing artificial reverberators," in *Proceedings of the 90th AES Convention*, 1991.

[43] J.-M. Jot, "Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces," *Multimedia Systems*, vol. 7, pp. 55–69, 1999.

[44] J.-M. Jot and J.-M. Trivi, "Scene description model and rendering engine for interactive virtual acoustics," in *AES 120th Convention*, 2006.

[45] M. A. Gerzon, "Unitary (energy preserving) multichannel networks with feedback," *Electronics Letters*, vol. 12, pp. 278–279, 1976.

[46] J. Stautner and M. Puckette, "Designing multichannel reverberators," *Computer Music Journal*, vol. 6, pp. 52–65, 1982.

[47] J. O. Smith III, *Physical Audio Signal Processing*. `http://www.dsprelated.com/dspbooks/pasp/History_FDNs_Artificial_Reverberation.html`.

[48] J.-M. Jot, "An analysis/synthesis approach to real-time artificial reverberation," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 2, pp. 221–224, 1992.

[49] J. O. Smith and D. Rocchesso, "Connections between feedback delay networks and waveguide networks for digital reverberation," in *Proceedings of the 1994 International Computer Music Conference*, 1995.

[50] E. Vickers, J.-L. L. Wu, P. G. Krishnan, and R. N. K. Sadanandam, "Frequency domain artificial reverberation using spectral magnitude decay," in *Proc. AES 121th Convention*, 2006.

[51] J. Blauert, *Spatial Hearing*. MIT Press, Cambridge, MA, 1997.

[52] Microsoft, "Introducing DirectX 9.0." `http://msdn.microsoft.com/en-us/library/bb318697(VS.85).aspx`, 2009.

[53] L. Savioja, J. Huopaniemi, T. Lokki, and R. Vaananen, "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc*, vol. 47, no. 9, pp. 675–705, 1999.

[54] J.-M. Jot, M. Walsh, and A. R. Philp, "3D audio renderer." United States Patent 20080037796.

[55] "The CIPIC HRTF Database." `http://interface.cipic.ucdavis.edu/CIL_html/CIL_HRTF_database.htm`.

[56] "LISTEN HRTF Database." `http://recherche.ircam.fr/equipes/salles/listen/`.

[57] S. Wilkinson, "Out of my head," *Electronic Musician*, vol. 5, 2001.

[58] J.-M. Jot, O. Warusfel, and V. Larcher, "Digital signal processing issues in the context of binaural and transaural stereophony," in *AES 98th Convention*, 1995.

[59] J. Huopaniemi, L. Savioja, and M. Karjalainen, "Modeling of reflections and air absorption in acoustical spaces a digital filter design approach," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.

[60] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. On Speech And Audio Processing*, vol. 6, p. 476, 1998.

[61] G. Theile, "Wave field synthesis a promising spatial audio rendering concept," in *Proc. of the 7th Int. Conference on Digital Audio Effects*, 2004.

[62] Wikipedia, "Wave field synthesis." http://en.wikipedia.org/wiki/Wave_field_synthesis.

[63] A. Devantier, S. Hess, and S. Olive, "Comparison of loudspeaker-room equalization preferences for multichannel, stereo, and mono reproductions: Are listeners more discriminating in mono?," in *AES 124th Convention*, 2008.

[64] T. Funkhouser, J.-M. Jot, and N. Tsingos, "'Sounds Good to Me!' computational sound for graphics, virtual reality, and interactive systems," in *SIGGRAPH 2002, Course Notes 45*, 2002.

[65] U. Kristiansen, A. Krokstad, and T. Follestad, "Extending the image method to higher-order reflections," *J. Applied Acoustics*, vol. 38, pp. 195–206, 1993.

[66] T. Funkhouser, N. Tsingos, I. Carlbom, G. Elko, M. Sondhi, J. E. West, G. Pingali, P. Min, and A. Ngan, "A beam tracing method for interactive architectural acoustics," *J. Acoust. Soc. Am.*, vol. 2, pp. 739–756, 2004.

[67] N. Tsingos, "Pre-computing geometry-based reverberation effects for games," in *AES 35th International Conference*, 2009.

[68] Y. Li, P. F. Driessen, and G. Tzanetakis, "Spatial sound rendering using measured room impulse responses," in *Proc. IEEE Int. Symposium on Signal Processing and Information Technology*, 2006.

[69] Z. Chen, *An Investigation of Acoustic Impulse Response Measurementand Modeling For Small Rooms*. PhD thesis, Montana State University, 2007.

[70] G. B. Stan, J. J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *J. Audio Eng. Soc.*, vol. 50, pp. 249–262, 2002.

[71] Wikipedia, "IBM Cell." http://en.wikipedia.org/wiki/Cell_microprocessor.

[72] A. Kemmler, "Acting on impulse," *Electronic Musician*, vol. 6, 2006.

[73] D. Miller, "Audio alchemy," *Electronic Musician*, vol. 6, 2008.

[74] Samplicity, "Samplicity impulse response library." http://www.samplicity.com/.

[75] "Open impulse response library." http://www.irlibrary.org/index.php.

[76] AudioEase, "Altiverb 6." http://www.audioease.com/Pages/Altiverb/AltiverbMain.html.

[77] Vienna Symphonic Library, "Vienna Multi-Impulse Response." http://vsl.co.at/en/211/497/1687/455/1714/1322.htm.

[78] J. H. Rindel, C. Lynge, G. Naylor, and K. Rishoj, "The use of a digital audio mainframe for room acoustical auralization," *96th Convention of the Audio Engineering Society*, 1994.

[79] S. Spors, H. Teutsch, and R. Rabenstein, "High-quality acoustic rendering with wave field synthesis," in *Vision, Modelling and Visualization*, 2001.

[80] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of wave field synthesis revisited," in *AES 124th Convention*, 2008.

[81] Prosoniq Products Software GmbH, "Rayverb technology whitepaper." http://www.prosoniq.com/whitepapers/rayverb-technology-whitepaper/.

[82] J. Edwards, "Acoustic room response analysis." TechOnLine Publication, 1997.

[83] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. 108th AES Convention*, 2000.

[84] P. Fausti, A. Farina, and R. Pompoli, "Measurements in opera houses: comparison between different techniques and equipment," in *Proc. of ICA98 - Int. Conf. on Acoustics*, 1998.

[85] I. Mateljan, "Signal selection for the room acoustics measurement," in *Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999.

[86] T. Collins, "A non-linear technique for room impulse response measurement," in *Proc. 6th Int. Conf. Digital Audio Effects (DAFX-03)*, 2003.

[87] R. Bristow-Johnson, "A little MLS (maximum-length sequence) tutorial." http://www.dspguru.com/dsp/tutorials/a-little-mls-tutorial.

[88] S. Muller and P. Massarani, "Transfer-function measurement with sweeps," *J. Audio Eng. Soc.*, vol. 49, pp. 443–471, 2001.

[89] J. Merimaa, "Applications of a 3-D microphone array," in *AES 112th Convention*, 2002.

[90] J. Pekonen, "Microphone techniques for spatial sound," in *Proceedings of the 2008 Acoustics Seminar on Spatial Sound Modeling* (M. Karjalainen, ed.), (Espoo, Finland), TKK Helsinki University of Technology, Department of Signal Processing and Acoustics, May 2008. Available online `http://www.acoustics.hut.fi/~jpekonen/Papers/`.

[91] Shure Incorporated, "Microphone techniques for studio recording." `http://www.shure.com/americas/support/publications/index.htm`.

[92] J. D. Johnston and Y. H. V. Lam, "Perceptual soundfield reconstruction," in *Proc. 109th AES Convention*, 2000.

[93] J. D. Johnston and E. R. Wagner, "Microphone array for preserving soundfield perceptual cues," January 2005.

[94] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Microphone array measurement system for analysis of directional and spatial variations of sound fields," *J. Acoust. Soc. Am.*, 2002.

[95] J. O. Smith III, *Introduction to Digital Filters: with Audio Applications*. W3K Publishing, 2007.

[96] ISO/FDIS 3382, *Acoustics - Measurement of the Reverberation Time of rooms with reference to other acoustical parameters*. Int. Organisation for Standardisation (1997), 1997.

[97] J. Huopaniemi, L. Savioja, and M. Karjalainen, "Modeling of reflections and air absorption in acoustical spaces: a digital filter design approach," in *Proc. 1997 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'97)*, 1997.

[98] J. Huopaniemi, *Virtual acoustics and 3-D sound in multimedia signal processing*. PhD thesis, Helsinki University of Technology, 1999.

[99] K. Steiglitz and L. McBride, "A technique for the identification of linear systems,," *IEEE Trans. Automatic Control, Vol. AC-10*, 1965.

[100] B. Friedlander and B. Porat, "The modified Yule-Walker method of ARMA spectral estimation," *IEEE Trans. on Aerospace Electronic Systems*, vol. AES-20, no. 3, 1984.

[101] L. Rabiner, J. McClellan, and T. Parks, "FIR digital filter design techniques using weighted chebyshev approximations," *Proc. IEEE*, vol. 63, 1975.

[102] H. Kuttruff, *Room Acoustics*. Elsevier Applied Science, London, UK, 1991.

[103] B. Gunel, "Room shape and size estimation using directional impulse response measurements," in *Proc. 3rd EAA Congress on Acoustics, Forum Acusticum Sevilla (CD-ROM)*, 2002.

[104] T. Lokki and M. Karjalainen, "An auditorily motivated analysis method for room impulse responses," in *Proc. of the COST G-6 Conf. on Digital Audio Effects*, 2000.

[105] D. Sunday, "Geometry algorithms." `http://softsurfer.com/Archive/algorithm_0104/algorithm_0104B.htm`.

[106] U. Zoler, *Digital Audio Effects*. John Wiley & Sons, 2002. `http://www.dafx.de/`.

[107] T. I. Laakso, V. Vlimki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay — tools for fractional delay filter design," *IEEE Signal Processing Mag.*, vol. 13, pp. 30–60, 1996.

[108] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, 2007.

[109] "ITU-R BS.1387: Method for objective measurements of perceived audio quality (PEAQ)."

[110] S. Bech, "Methods for subjective evaluation of spatial characteristics of sound," in *16th AES International Conference: Spatial Sound Reproduction*, 1999.

[111] J. Berg and F. Rumsey, "Systematic evaluation of perceived spatial quality," in *24th AES International Conference: Multichannel Audio, The New Reality*, 2003.

[112] "ITU-R BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems (MUSHRA)."

[113] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application*. John Wiley & Sons, Ltd, 2006.

[114] M. Schonle, U. Zolzer, and N. Fliege, "Modeling of room impulse responses by multirate systems," in *Proc. 93rd AES Convention*, 1992.

[115] Y. Haneda, S. Makino, and Y. Kaneda, "Common acoustical pole and zero modeling of room transfer functions," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 320 – 328, 1994.