Spatial auditory display for acoustics and music collections

Stewart, Rebecca

Spatial Auditory Display for Acoustics and Music Collections

Rebecca Stewart

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of the
University of London.

School of Electronic Engineering and Computer Science
Queen Mary, University of London

July 2010

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline. I acknowledge the helpful guidance and support of my supervisor, Prof Mark Sandler.

Abstract

This thesis explores how audio can be better incorporated into how people access information and does so by developing approaches for creating three-dimensional audio environments with low processing demands. This is done by investigating three research questions.

Mobile applications have processor and memory requirements that restrict the number of concurrent static or moving sound sources that can be rendered with binaural audio. Is there a more efficient approach that is as perceptually accurate as the traditional method? This thesis concludes that virtual Ambisonics is an efficient and accurate means to render a binaural auditory display consisting of noise signals placed on the horizontal plane without head tracking. Virtual Ambisonics is then more efficient than convolution of HRTFs if more than two sound sources are concurrently rendered or if movement of the sources or head tracking is implemented.

Complex acoustics models require significant amounts of memory and processing. If the memory and processor loads for a model are too large for a particular device, that model cannot be interactive in real-time. What steps can be taken to allow a complex room model to be interactive by using less memory and decreasing the computational load? This thesis presents a new reverberation model based on hybrid reverberation which uses a collection of B-format IRs. A new metric for determining the mixing time of a room is developed and interpolation between early reflections is investigated. Though hybrid reverberation typically uses a recursive filter such as a FDN for the late reverberation, an average late reverberation tail is instead synthesised for convolution reverberation.

Commercial interfaces for music search and discovery use little aural information even though the information being sought is audio. How can audio be used in interfaces for music search and discovery? This thesis looks at 20 interfaces and determines that several themes emerge from past interfaces. These include using a two or three-dimensional space to explore a music collection, allowing concurrent playback of multiple sources, and tools such as auras to control how much information is presented. A new interface, the amblr, is developed because virtual two-dimensional spaces populated by music have been a common approach, but not yet a perfected one. The amblr is also interpreted as an art installation which was visited by approximately 1000 people over 5 days. The installation maps the virtual space created by the amblr to a physical space.

Acknowledgements

I would first like to thank my supervisor, Mark Sandler. This thesis would not have been possible without his support and guidance. I would also like to thank the various funding bodies that have allowed me to conduct this research: Queen Mary, University of London; the Audio Engineering Society Education Foundation; the University of London Central Research Fund; and the Google Anita Borg Memorial Scholarship.

A wise IRC conversation once said "it takes a whole research group to write a thesis." While that same conversation also instructed not to hold a thesis by its legs, at least the part about the research group was right. This thesis would not have been possible or half as much fun without the people that make up C4DM. In particular: Adam, Alice, Amélie, Andrew R., Andrew N., Chris C., Chris L., Chris S., Dan, Enrique, Ivan, Jean-Baptiste, Katy, Kurt, Louis, Louise, Magdalena, Maria, Martin, Matthew, Matthias, Michael, Robin, Thomas, Xue, and Yves. I also would like to thank all my listening test and user evaluation participants.

Finally, I would like to thank my family. They may not have always understood what I was doing, but they have always supported me regardless. I am especially grateful to the Editor. Ben, thank you. Words cannot express how much your support has meant to me. I only hope I can return the favor.

Contents

List of Figures

List of Tables

List of Symbols

$\otimes$	linear convolution
η	echo density profile
γ_4	kurtosis
ϕ	elevation
τ	delay preceding the direct sound
θ	azimuth
A	absorption coefficient
C_N	clarity where N is 50 or 80
f_c	Schroeder frequency
f_s	Sampling frequency
g	gain
h	impulse response
l	loudspeaker vector
p	phantom sound source vector
$RT_{N,f}$	reverberation time
r_e	Gerzon's localisation energy vector
r_v	Gerzon's localisation velocity vector
S	surface area
s	sine sweep
t_c	Schroeder diffuse-time interval
t_{mixing}	mixing time
t_{thresh}	limit in time after which direct sound when mixing time can occur
V	volume
x	audio input
y	audio output

List of Abbreviations

ANOVA	analysis of variance
CAVE	cave automatic virtual environment
C4DM	Centre for Digital Music
DWM	digital waveguide mesh
EDC	energy decay relief
FDN	feedback delay network
FFT	Fast Fourier Transform
GUI	graphical user interface
HCI	human-computer interaction
HRTF	head-related transfer function
IACC	interaural cross-correlation coefficient
IFFT	inverse fast Fourier transform
ILD	interaural level difference
IR	impulse response
ITD	interaural time difference
LEV	listener envelopment
LTI	linear time-invariant
MIR	music information retrieval
MIT	Massachusetts Institute of Technology
MLS	maximum-length-sequence
MSE	mean square error
STFT	short-time Fourier transform
VBAP	vector-based amplitude panning
WFS	wave field synthesis

Chapter 1

Introduction

Weekday commutes in London produce buses and train carriages full of people immersed in mobile devices heading to or from offices. Music players blast the white earbud-adorned masses while they furiously type text messages or scour the internet for news or entertainment. In central London, the tourist industry is also hard at work as millions of visitors flood the city's attractions. These visitors, too, are clinging to mobile devices. GPS-enabled smartphones guide them through the streets and towards where they can capture and tweet their own iconic photo of Big Ben. In the British Museum, many of the 600 million annual visitors use handheld devices to view multimedia presentations about 200 of the objects from the vast collections. Across the river, visitors to the Tate Modern view the galleries while holding audio guides which give additional insights into the artwork.

Electronics and computing are becoming increasingly mobile and pervasive throughout day-to-day activities. These devices are used for entertainment, education, and communication. As mobile devices cannot support the screen size of a desktop computer, they are forcing new paradigms of interaction that cannot rely solely on visual information. Non-visual interaction was once designated as assistive technology, somehow believed that when given an option, people would prefer to interact with an object visually over other modes. Humans are multi-modal and those different modes have evolved for particular tasks. There are tasks that are best completed with visual interaction, but there are also tasks that are better suited to audio. This thesis takes the viewpoint that more interfaces would use spatial audio if there was a means to easily incorporate it into mobile devices.

Headphones are a widespread means of escape. It may be an escape from the cacophony of a commute into a favourite playlist or from masses of fellow tourists into a personal and private exploration of history. Since headphones are commonly worn and usually expected when using a mobile device, why not use more audio to complete the task at hand? Spatial audio can enhance the interaction and create a more immersive experience. However, to incorporate more audio into handheld devices, there needs to be a simple means to generate and manipulate that audio. This thesis defines a preliminary list of requirements for a spatial audio interface to be:

Mobility Audio is played over headphones; can be run on a mobile device; no additional hardware beyond the device and headphones is required.

Flexibility Accommodate a range of designs and applications; generic software architecture that could be implemented in a number of ways from mobile devices to web services.

Accuracy Be perceptually convincing; sound sources feel as if they occupy a specific point in space; the same sound source needs to appear to be in the same location for multiple people; it cannot be different for different users.

This thesis looks at spatial auditory display from a signal processing perspective. Spatial auditory display is the presentation of information with virtual spatial audio environments. When creating a spatial auditory display for a mobile device, it is important to consider the signal processing demands and restraints. This thesis develops a signal processing platform for spatial audio display and then demonstrates its abilities through the development of two different interfaces.

1.1 Research Question

The motivating question for this thesis is: *How can spatial audio be better incorporated into how people access information?* This thesis approaches this primary question by further dividing it into three sub-questions:

1. Mobile applications have processor and memory requirements that restrict the number of concurrent static or moving sound sources that can be rendered with binaural audio. Is there a more efficient approach that is as perceptually accurate as the traditional method? Chapter 3 addresses this question.

2. Complex acoustics models require significant amounts of memory and processing. If the memory and processor loads for a model are too large for a particular device, that model cannot be interactive in real-time. What steps can be taken to allow a complex room model to be interactive by using less memory and decreasing the computational load? Chapter investigates these questions.

3. Commercial interfaces for music search and discovery use little aural information even though the information being sought is audio. How can audio be used in interfaces for music search and discovery? Chapter 5 surveys the literature for previous solutions and Chapter 6 builds on these solutions with experimental studies.

1.2 Thesis Outline

Chapter 1 – Introduction establishes the motivation and objectives of the thesis. The major contributions and publications are listed.

Chapter 2 – Spatial Hearing and Sound Field Reproduction introduces the basic technical concepts needed to understand the subsequent chapters. It begins with an overview of how humans localise single sources in a free field, that is without reverberation. The chapter then moves on to how reverberation is perceived.

The model of direct sound, early reflections, and late reverberation is introduced, and the psychoacoustic properties of early reflections and late reverberation are discussed in greater detail. An overview of how a spatial sound field is modelled and played for a listener is then surveyed including statistical and physical models for reverberation. The methods for conveying point sources and reverberant sound fields via loudspeakers and then headphones are reviewed.

Chapter 3 – Optimising Binaural Auditory Display introduces the concept of auditory display and the requirements for binaural auditory displays. A number of common application areas and the design features often employed are outlined. The chapter then establishes the goals and restrictions of presenting information via binaural audio while touching on the history of the field and movement towards mobile, eyes-free computing.

The psychoacoustical and signal processing constraints on creating interactive spatial auditory displays are summarised and virtual Ambisonics is proposed as a flexible, efficient signal processing platform. An optimisation of virtual Ambisonics is derived for higher orders and ideal parameters for the system are reviewed. Virtual Ambisonics is evaluated to observe the effect of decoding order and virtual loudspeaker configuration on the binaural signal. Signal analysis shows measurable differences between the virtual Ambisonics approaches and measured HRTFs, but listening tests show that these differences are not perceptible.

Chapter 4 – A Novel Reverberation Model for Real-Time Auralization presents auralization of room models as spatial auditory displays. Motivation for room auralization is covered and the requirements for a system to qualify as real-time are discussed. Hybrid reverberation is reviewed, emphasising the deterministic-stochastic model of an impulse response. A new method for automatically extracting the mixing time of a room from an impulse response is presented. The method is used to extract the early reflections from multiple impulse responses so that new early reflections can be interpolated. The mixing time and interpolation methods are then evaluated. The mixing time method is also applied to determine the late reverberation of an impulse response. An averaged late reverberation tail representative of an entire room is modelled so that simulation of movement in an auralization model can occur without updating every filter coefficient of the impulse response. The resulting reverberation is analysed with acoustic metrics.

Chapter 5 – Music Collection Auditory Displays introduces music information retrieval in the context of music exploration and discovery. It explores the design history and patterns of audio within music informatics interfaces, highlighting areas of improvement. Common design paradigms are discussed, particularly the use of two-dimensional maps of musical spaces. The spatialisation algorithms for the interfaces are reviewed.

Chapter 6 – The amblr:A Novel Music Browser documents the design and evaluation of the amblr, an interface for exploring a collection of music. The amblr is developed in two iterations and is evaluated by users after the first iteration. The amblr builds on established approaches from previous work and pulls together disparate auditory display tools while incorporating additional novel design features. It explores a collection of music by allowing a user to navigate through a collection without relying on text. The driving concept is that for effective music exploration and discovery, audio content should be presented before metadata.

The amblr navigates a two-dimensional arrangement of music. All of the previous interfaces that use a similar approach to navigate a collection of music rely heavily on visual displays to accompany auditory displays. The amblr has a visual display, but it is minimal; the auditory display is the primary mode of interaction. Although the interface is fully functional without any visuals, the graphical user interface aids interaction by visually illustrating the different parameters of the interface. The amblr takes further advantage of computing on handheld devices by using gestural controls and haptic feedback driven by integrated sensors such as accelerometers.

Chapter 7 – Conclusions and Future Work summarises the findings of the thesis and outlines future research.

1.3 Major Contributions

This thesis contributes:

Chapter 3 – Extension of optimised virtual Ambisonics to second and third order; study of the effects of virtual loudspeaker placement and decoding order on a binaural signal.

Chapter 4 – Novel model for interacting with a large collection of impulse responses in order to simulate real-time movement; novel approach to determining the mixing time of a room; study of interpolating early reflections of room impulse responses from differing positions and distances relative to the interpolated impulse response; novel approach to averaging the late reverberation of multiple room impulse responses.

Chapter 5 – First review and analysis of how auditory display has been applied to music information retrieval tasks.

Chapter 6 – Novel interface for exploring a collection of music with gestural control without visual feedback; a set of novel navigation tools for exploring a collection of music; a unique combination of established and new tools for exploring a collection of music; novel art installation and physical interface for exploring a collection of music.

Appendix B – Largest known publicly-available database of omnidirectional and B-format room impulse responses.

1.4 Publications

Conference Proceedings

R. Stewart and M. Sandler, “Real-time panning convolution reverberation.” In Proc. of 123rd AES Conv., New York, USA, October 2007.

R. Stewart and M. Sandler, “Statistical measures of early reflections of room impulse responses.” In Proc. of the 10th Int. Conf. on Digital Audio Effects (DAFx-07), Bordeaux, France, September 2007.

R. Stewart and M. Sandler, “3D interactive environment for music collection navigation.” In Proc. of the 11th Int. Conf. on Digital Audio Effects (DAFx-08), Helsinki, Finland, September 2008.

R. Stewart and M. Sandler, “Generating a spatial average reverberation tail across multiple impulse responses.” In Proc. of the 35th Int. AES Conf. on Audio for Games, London, UK, February 2009.

R. Stewart and M. Sandler, “Database of omnidirectional and B-format room impulse responses.” In Proc. of IEEE Conf. on Acoustics, Speech, and Signal Processing (ICASSP ’10), Dallas, Texas, USA March 2010.

Peer-Reviewed Installations and Posters

M. Magas, R. Stewart, and B. Fields, *decibel151*, Shown at Information Aesthetics at ACM SIGGRAPH, New Orleans, USA, August 2009.

R. Stewart, S. Lloyd, M. Sandler, and M. Magas, “Scalable spatial audio browser,” Shown at Int. Society for Music Information Retrieval Conf. (ISMIR’09). Kobe, Japan, October 2009

R. Stewart, “The amblr: a new way to explore a music collection.” Presented at the Grace Hopper Celebration for Women in Computing 2010, Atlanta, Georgia, USA September 2010.

Peer-Reviewed Journal Articles

(white paper accepted and full paper submitted) R. Stewart and M. Sandler, Interactive Auditory Display for Mobile Music. *IEEE Signal Processing Magazine Special Issue on Multimedia Search,* July 2011.

(submitted) R. Stewart and M. Sandler, Optimization and Perception of Virtual Ambisonics. *IEEE Trans. on Audio, Speech and Language Processing.*

(submitted) R. Stewart and M. Sandler, Spatial Auditory Display in Music Search and Browsing Applications. *Int. J. of Human Computer Studies.*

Patent Applications

M. Sandler and R. Stewart, "Method and Apparatus for Music Collection Navigation", UK Patent Application No. GB0815362.9, filed September 2009.

Chapter 2

Spatial Hearing and Sound Field Reproduction

An understanding of how sound is perceived and how that perception can be manipulated is necessary to discuss the design of spatial auditory display. This chapter will review the foundations of how humans perceive sound, in particular localising and interpreting sound in a spatial sound field. The perception of a single sound source without any room information such as reverberation is first covered. Controlled laboratory experiments differ greatly from realistic listening situations as people seldom encounter sound without an environmental context in day-to-day life. Even so, such experiments have established a basic model of how humans localise sounds which is presented in Section 2.1. Section 2.2 examines how environmental sounds such as reverberation are perceived. The most popular techniques for creating artificial spatial environments are then reviewed. Section 2.3 looks at models that have been developed to simulate a room, then Sections 2.4 and 2.5 study how those models can be rendered for a listening environment. This includes both loudspeaker and headphone presentation along with a discussion of how to choose the most appropriate system for a given task.

2.1 Localisation of Single Sources

Researchers study the localisation of single sound sources without any environmental noise by creating artificial listening scenarios using headphones and loudspeakers. From this, a basic theory of how humans localise sound has been developed. It is acknowledged that humans respond differently to the complex sound scenes that exist in the natural world than in the simplified artificial scenes created in a laboratory environment. However, these artificial listening environments have shown that manipulating fundamental cues based on level, timing, and frequency information is the basis for synthesising spatial sound fields.

For clarity, the direction of a sound source will always be referenced relative to the listener's head; sound sources will be placed with references to a set of planes which intersect the listener's head. The three planes are illustrated in Figure 2.1 and are defined as: the horizontal plane parallel to the shoulders and dividing what is above and below the listener's ears; the frontal plane dividing what is in front of and behind the listener; and the median plane which divides what is to the left and right of the listener. A sound

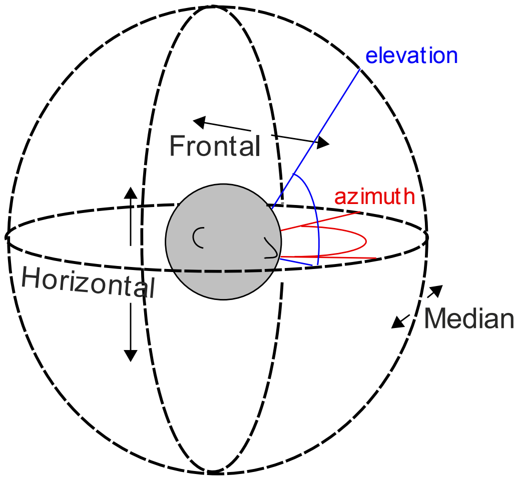


Figure 2.1: Definition of the three planes centred around the listener's head.

source location will also be described in terms of azimuth and elevation. A source with an azimuth of 0° lies on the median plane whilst a source with an elevation of 0° lies on the horizontal plane. An azimuth of 0° is directly in front of the listener with an azimuth of 90° to the right, −90° to the left, and 180° directly behind the listener's head.

Experimental work in spatial audio refers to lateralisation as well as localisation. Lateralisation refers to the perceived position of a sound source 'in the head' between the two ears as can occur with headphone listening. The term localisation implies that the sound source is either a real sound source, is presented with loudspeakers, or appears to be located outside the head when listening to headphones [Moore, 2003, p. 233].

2.1.1 Physiology

While much of human hearing is reliant on the middle and inner ear, it is the outer ear that most influences spatial hearing. The outer ear consists of the pinna and the ear canal illustrated in Figure 2.2. The pinna is the portion of the outer ear that is outside the head. Everyone has a unique pinna, like a fingerprint. It was once believed that the pinna served no function other than to "funnel" sound and to protect the middle and inner ear, but this has been found to be untrue. The pinna plays a large role in spatial hearing and also has been shown to decrease wind noise [Blauert, 1997, p. 53].

On average, the pinna is angled 25° to 45° from the head [Blauert, 1997, p. 53]. The pinna surrounds the entrance of the ear canal, the tube leading to the ear drum. The sound is transmitted to the middle ear via the ear drum. The ear canal has no propagation loss so all resonances and filtering effects from the outer ear are transmitted without any attenuation [Blauert, 1997, p. 56]. The ear canal resonates at approximately 4 kHz; this increased sensitivity to 4 kHz frequencies can be seen in equal-loudness curves [Zwicker and Fastl, 1999, p. 24].

The head, shoulders, and outer ear encode spatial information about a sound source into temporal and spectral attributes, though it is not completely understood how this is done [Blauert, 1997, p. 63]. It is known that the pinna acts as a linear filter at

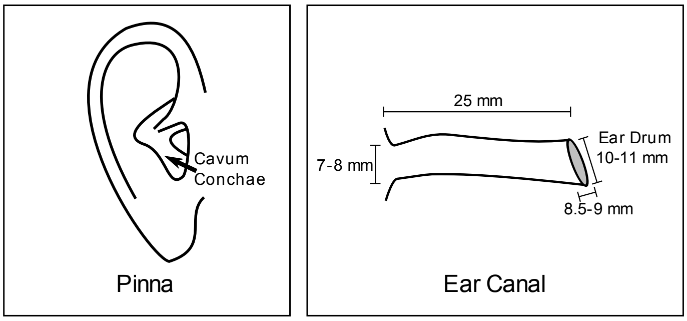


Figure 2.2: The outer ear.

higher frequencies, but a general model has not been developed. This is partly due to the uniqueness of individual pinnae. The skin and cartilage that form the pinna reflect and diffract sound. Further filtering occurs from the shoulders and head which affect frequencies below 1.5 kHz [Zwicker and Fastl, 1999, p. 24].

2.1.2 Duplex Theory

Two of the basic cues for sound localisation are the frequency-dependent interaural time difference (ITD) and interaural level difference (ILD). Together, they are referred to as Duplex Theory which was developed by Lord Rayleigh at the turn of the 20th century [Rayleigh, 1907].

ITD is the time or phase difference caused by the distance the sound travels from the source to each ear. It is the main cue for frequencies below about 1 kHz. Early researchers in spatial audio perception believed that ITD is the most important cue (for example see Rayleigh [1907]), but Blauert [1997, p. 141] strongly cautions against over-generalising the results of very simplified listening environments to 'natural listening.'

The ears are on average 18 cm apart, so unless a sound source is directly on the median plane (the plane that encompasses locations directly in front, above, or behind the listener) the sound will take longer to travel to one ear than the other [Howard and Angus, 2006, p. 97]. ITD applies primarily to lower frequencies because it is difficult to precisely detect differences between the two ears for wavelengths significantly smaller than the human head. However, higher frequencies can be localised with other cues.

ILD is the difference between the loudness or intensity of the sound signal at each ear due to head-shadowing. Wavelengths smaller than the head, e.g. frequencies above about 1.5 kHz, are localised using ILD. Longer wavelengths diffract around the head, and frequency-content below 500 Hz has negligible ILD. However, frequency content below 500 Hz may be effected by ILD if the sound source is close to the head [Moore, 2003, p. 236]. The difference between path lengths of differing source positions is illustrated in Figure 2.3.

Although ITD is generally used for frequencies below 1 kHz and ILD is used for frequencies above 1.5 kHz, these are not abrupt nor absolute transitions. ITD and ILD are complementary cues and can be used at the same time, though one cue can override the other. For example, if a is click presented to a listener via headphones at the same time and level to each ear, it appears to be centred. If the click in the left ear precedes

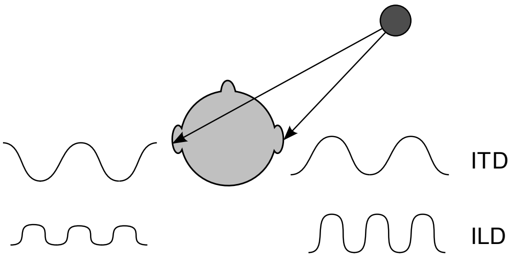


Figure 2.3: When a sound source is not on the media plane, there are differences in the sound that arrives at each ear. ITD is the difference in timing or phase and ILD is a difference in loudness.

the click in the right ear by 100 μs, the click appears to be shifted left. If the click in the right is then increased in volume, the perceived location returns to the centre. This is referred to as time-intensity trading [Moore, 2003, p. 244-245].

2.1.3 Frequency cues

Humans need more information than is provided by Duplex Theory cues in order to discriminate between ambiguous source locations along the median plane and the cone of confusion as illustrated in Figure 2.4. The cone of confusion is a region extending from one ear where every location on the surface of the cone has similar ITD and ILD cues. Similar cues also occur along the median plane as the sound source is equidistant from both ears. To further determine where a sound is located along the median plane or a cone of confusion, additional information is available in the frequency content of the signals at the ears.

The audio that reaches a human's inner ears undergoes a series of delays and reflections before reaching the ear drum. While ILD and ITD cues largely influence the perceived location of a sound source, when these cues alone are used to produce a stereo signal over headphones, the sound source can appear to be between the listener's ears. By processing the sound source with the delays and filters usually introduced from reflections from the ears, head, and torso, the source can appear to be outside the listener's head. High frequencies above 6 kHz are particularly important as their wavelengths are short enough to reflect off the pinna, but lower frequencies are also informative as they bounce off of the head and torso [Moore, 2003, p. 251].

2.1.4 Precedence effect

The world is a noisy environment in which humans constantly have spatial audio streams to process. Humans have learned to parse the audio environments around them using a complex array of auditory cognition processes which are of particular importance to binaural auditory display.

The precedence effect goes by many names and includes a variety of caveats. Some refer to it as the Haas effect after Haas [1972]. Blauert [1997] refers to 'summing localisation' and the 'law of the first wavefront' as two elements of the precedence effect. The most basic interpretation is that if two coherent but delayed signals are presented

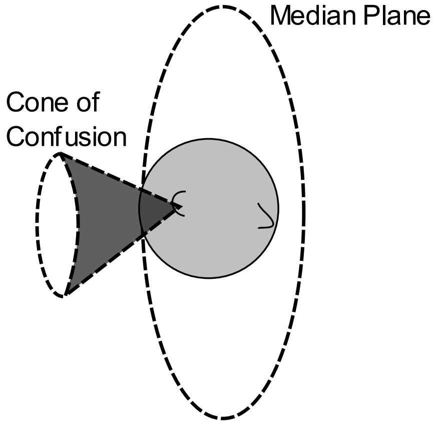


Figure 2.4: Illustration of where Duplex Theory cues are insufficient for disambiguation of source location: the median plane and cone of confusion.

to a listener the first sound to arrive is heard but the second is suppressed. This leads the listener to localise the sound source to the perceived location of the first sound.

Summing localisation applies to coherent signals that are offset by no more than 1 ms from each other. Multiple sound sources that reach a listener within 1 ms appear as a fused single source that is located as a compromise between the original sources. That compromised location is determined by a variety of factors including the content of the signal [Blauert, 1997, p. 215]. Summing localisation is demonstrated with commercial stereo recordings as sound sources appear as 'phantom images' that emanate between the loudspeakers. Summing localisation also ruins the intended spatial sound field if a listener is too close to a single loudspeaker as the majority of the sound sources will appear to only be emanating from that loudspeaker. The two signals are perceived as one and are shifted towards the location of the earlier signal as the delay increases.

When the time difference is greater than 1 ms, the perceived position remains constant and is determined by the law of the first wavefront [Blauert, 1997, p. 222]. The principle of the law of the first wavefront is that the first wavefront to arrive at the listener dictates the direction of the source because delayed versions (reflections off of surrounding surfaces) are psychoacoustically suppressed [Haas, 1972]. However, even though the subsequent reflections from varying directions seem to be ignored, the spatial information from the reflections is not lost, and listeners do notice a difference between the inclusion of echoes and their absence[Moore, 2003, p. 255].

The aspect of the precedence effect addressing echo suppression is also called the Haas effect. Echo suppression can be substantial: at a delay of 32 ms a reflection can be 5 dB louder than the initial sound and be perceived as inaudible; a delay less than 15 ms can be 10 dB louder without being perceived; and in general reflections less than 50 ms are not perceived as 'annoying' [Blauert, 1997, p. 226].

The precedence effect has been found to be event-driven and may break down when a listener is presented with information that appears 'implausible,' such as a source jumping from one location to another unexpectedly. This can increase the amount of time a listener needs to become accustomed to a sound field before being able to localise sounds [Clifton, 1987]. The content of the signal also influences the perceived location and whether the precedence effect is applied. The two signals need to be similar to each other, and the type of signal will determine the maximum delay before two separate sources are perceived. Single clicks need to be less than 5 ms apart to appear as a single sources whilst music or speech signals can be delayed by up to 40 ms [Moore, 2003, p. 253].

2.1.5 Cocktail party effect

When multiple audio objects are recognised by the listener as separate sound objects, further neural processing allows the listener to understand the entire sound field and to shift focus or attention between those objects. The cocktail party effect is a phrase that was first coined by Cherry [1953] and describes selectively focusing on a single sound source in the presence of competing sounds. Focusing and understanding a single sound source amongst a cacophony of other noises is dependent on spatial separation of the sounds. Cherry [1953] conducted the first formal study on this effect and found that when two speaking voices were presented equally to both ears as a mono signal, listeners could not identify what was said. When the voices were spatially separated and only one voice was in each ear of the headphones, the listeners had no problems reiterating the content of one of the two presented speech signals.

Research has shown that the more different each concurrent stream of audio is from the others, the better a listener can segregate the streams. Also, if a single stream is more predictable, a listener can better follow or focus on that stream [Bregman, 1999, p. 529]. A stream can be predictable if it is semantically connected, such as series of spoken words that make grammatical sense, or if it is statistically continuous, such as a sinusoidal signal [Arons, 1992].

A large proportion of research surrounding the cocktail party effect has concentrated on speech signals, but some work has examined the audio and music domain. [Brazil et al., 2009] found that listeners can listen to 3 to 6 concurrent audio clips of non-speech real-world sounds and correctly identify the sound sources. The sounds were presented monaurally via headphones, so there was no spatial separation, but sounds were offset by 300 ms. When the sounds were semantically linked – the sounds had a similar timbre or would naturally occur together in the real world – listeners correctly identified a greater percentage of the sounds. Lorho et al. [2001] played concurrent recordings of musical instruments. They found that offsetting the onsets and placing the sources in different spatial locations resulted in more accurate identification of sounds than concurrent sounds with identical onsets.

2.1.6 Perceptual limits/resolutions

While minute differences in signals can be detected and measured, the human auditory system is not consistently robust nor precise. The resolution of the auditory system varies according to temporal, spectral, and spatial factors of the presented sound field along with physiology and psychology of the listener.

Type of Signal	*Localisation Blur*
Impulses	$0.75° - 2°$
Impulse train	1.6°
Sinusoids	4.4°
Sinusoids	$> 1°$
Sinusoids	$1.1° - 4.0°$
Sinusoids	$1.0° - 3.1°$
Narrow-band noise, cos^2 tone bursts	$1.4° - 2.8°$
Gaussian tone bursts	$0.8° - 3.3°$
Speech	0.9°
Tone bursts with differing onset, decay times, and frequencies	$1.8° - 11.8°$
Speech	1.5°
Broadband noise	3.2°

Table 2.1: Survey of measurements of localisation blur for horizontal displacement of sound source away from the forward direction. Adapted from [Blauert, 1997, p. 39]

Localisation blur is the perceptual error in the human auditory system when determining the location of a sound source. It is usually measured as the minimum audible angle or just-noticeable difference of a sound source presented at different azimuths and elevations. In general, the localisation blur is smallest directly forward on the horizontal plane and increases towards the right and left ears between three and ten times the blur straight ahead. Behind the listener the localisation blur decreases again to about two times the forward value [Blauert, 1997, p. 40-41]. Localisation in the median plane is better directly in front of and behind the listener than directly above.

Like all localisation tasks, the localisation blur depends on the type of sound source. A survey of localisation studies signal types and resulting blur can be seen in Table 2.1. Broader band signals are generally more accurately localised. Signals with less than two-thirds of an octave of bandwidth cannot be correctly localised on the median plane and the perceived location is completely dependent on the frequency content [Blauert, 1997, p. 45]. However, familiarity with the sound source improves localisation, particularly on the median plane. Localisation of speech by an unfamiliar person is localised with a blur of 17° while speech by a familiar person is localised to within 9° [Blauert, 1997, p. 44].

The auditory system not only has difficulties with precise locations of static sources; it has difficulties noticing when a source is moving. The slower a sound source moves, the smaller the minimum audible movement angle. It has been found that when a source moves at a rate of 90°/s it needs to move 21° before the movement is noticed [Moore, 2003, p. 262].

Humans instinctively try to compensate for spatial hearing deficiencies by altering the cues received. Head movements may be conscious or unconscious, but it has been shown experimentally that most people move their heads when trying to resolve the location of a sound source [Blauert, 1997, p. 177-191]. Most first try to turn towards the sound, though not necessarily directly facing it, and then they move their heads up or down. This helps to compensate for poor spatial resolution behind, above, and below a listener by moving the sound source to the front. These movements also disambiguate cues from a source on the median plane or its location on a cone of confusion.

2.2 Perception of Reverberant Spaces

The angle and elevation of a sound source determine cues such as ITD and ILD, but angle and elevation are only a part of sound source localisation; distance completes the soundscape. If a listener is unfamiliar with a sound source, then loudness is often the overriding distance cue [Begault, 2000, p. 70]. Without considering environmental surroundings and acoustics such as reverberation, the loudness of a sound source is inversely proportional to its distance. However, if a listener is familiar with a sound, then the intensity needs to be aligned with the context. For instance, whispering and shouting have associated distances; the whispering sound source is interpreted as being closer to the listener than a shouting source regardless of loudness levels.

Whether a listener is or is not familiar with a sound source, other cues do greatly influence distance perception. One cue is an attenuation of high frequencies the farther a sound source is located. This is due to air absorption which is dependent on the humidity of the air [Begault, 2000, p. 77-78]. Also, the more reverberation present in a signal compared with the direct sound, the farther the source is perceived. Adjusting the direct to reverberation ratio cannot push a sound source to an infinite distance, though. Instead the sound will stop at what is called the auditory horizon and will not increase in perceived distance. The auditory horizon is dependent on many factors including characteristics of the reverberant signal. It places a perceptual limit on how far away a sound source can be placed in a virtual space [Sheeline, 1982]. The physics of reverberation will be discussed later in Section 2.2.3.

Reverberation may give a sense of depth or distance to a sound field, but it has also been found that it can make a sound source more difficult to localise and may make a sound source appear to be wider [Blauert, 1997]. This is often desired in musical performance spaces and may or may not be a hindrance in virtual audio environments. Depending on the purpose of the virtual environment, reverberation may have greater benefits than drawbacks as it has a significant impact on externalising audio presented over headphones [Begault, 2000, p. 80-82].

Most acoustic metrics are derived from a measurement called a room impulse response (IR). An IR is an audio recording of the reverberation in a space. It can be analysed to determine characteristics about the room such as the amount of time required for a sound to completely decay (reverberation time) or how easy it is to understand speech or musical performance (clarity). The specifics of how an IR is measured are covered in Section 2.3.1.

Reverberation consists of three basic components: the direct sound, early reflections, and late reverberation. When nothing is obstructing the sound source from the listener, the direct sound is the first wavefront to arrive. When the listener is familiar with

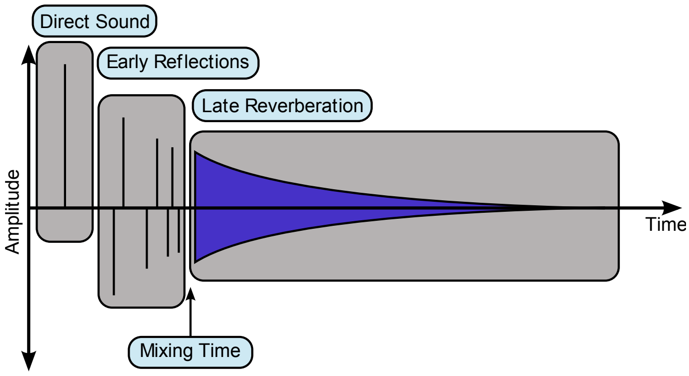


Figure 2.5: The three basic components of reverberation.

the sound source, the direct sound gives the first impression of distance according to its frequency characteristics [Blesser, 2001]. Following the direct sound are the early reflections, a set of discrete reflections whose density increases until individual reflections can no longer be discriminated and/or perceived. These early reflections are not perceived as individual echoes because of the precedence effect. When the early reflections are sufficiently dense and distributed throughout the space, they are considered to be late reverberation and the room is considered mixed.

2.2.1 Direct sound

The level of the direct sound influences the perceived distance of the source from the listener in a number of ways. As has been discussed, the familiarity and loudness of the event give an initial cue to the distance, but the amount of reverberant energy present after echo suppression from the precedence effect (approximately 40 ms) also indicates distance. If another discrete reflection arrives after the offset of the precedence effect, it is perceived as an echo, but if a significant level of late reverberation is still present, the sound source's location can no longer be distinguished. If the level of the direct sound and early reflections are low enough in comparison with the late reverberation, the late reverberation backward-masks the earlier sound and only reverberant energy is perceived. It has been found that if the diffuse reverberant energy is +3 dB higher than the direct sound, then the direct sound is not heard [Blauert, 1997, p. 279].

2.2.2 Early reflections

Figure 2.5 is a simplified illustration of the three basic components of reverberation. Though it shows early reflections to have a definite start and end, in reality it difficult to determine when the early reflections transition to late reverberation. If there is nothing obstructing the direct sound from the listening positions, then the first early reflection is the second sound to reach the listener after the direct sound. Mathematically, the early reflections are modelled as a deterministic function while the late reverberation is modelled as a stochastic function [Blesser, 2001]. They are also defined by the dimensions of a room. The number of reflections that arrive at a single point within the room grow exponentially with time:

$$\frac{dN_r}{dt} = 4\pi\frac{c^3t^2}{V} \tag{2.1}$$

where N_r is the number of reflections, t is the time from the direct sound, c is the speed of sound, and V is the volume of the room [Kuttruff, 2000].

Perceptually, the early reflections are significant as they are dependent on the location of the sound source, the listener, and the surrounding environment. As will be discussed further in Section 2.2.5, the energy contained in the early reflections can greatly improve or hinder understanding of speech. Lateral reflections enhance the listener envelopment (LEV) and apparent source width – how immersed in a sound field a listener feels and how wide a source is perceived to be [Gardner, 1998, p. 93], [Beranek, 1992].

Conventionally, the first 80 ms after the direct sound is deemed early reflections, but it is accepted that this is a fairly arbitrary value [Begault, 2000, p.83]. Further definitions of when early reflections end and the late reverberation begins are discussed in Section 2.2.4.

2.2.3 Late reverberation

The late reverberation strongly influences the time and frequency characteristics of sound sources as it comprises the majority of the IR. Blesser and Salter [2007] have an elegant analogy to describe reverberation.

> Think of a cathedral as millions of bells (resonating oscillators), each with its own pitch (resonance frequency), and each with a slightly different decay rate (reverberation time). The clarinet sound rings (excites) only those bells with a pitch corresponding to the frequency content of the clarinet. In other words, you are actually hearing the bells of space, not the original clarinet sound. With hundreds of millions of bells at all possible pitches, a cathedral can faithfully reproduce the frequency content of any musical instrument. [Blesser and Salter, 2007, p. 247]

If every room has its own set of 'bells', then one of the defining characteristics of that room is how long it takes each bell to return to silence after it has been excited. The same excitation signal can be presented to several rooms, but the acoustic properties of each individual space determine the frequency-dependent reverberation time and timbre of the space. LEV is influenced by the frequency-dependent reverberation time as well. Morimoto et al. [2007] found that shorter reverberation times at low or high frequencies decreases the LEV while lengthening increases LEV. So the envelopment of a space is dependent on the specific balance of frequencies.

2.2.4 Transition from early to late

Late reverberation is a noise-like signal that can be easily approximated with noise. Moorer [1979] first proved this by convolving noise with a sound source. The result was a reverberant signal that did not sound like a particular space. The noise-like signal of late reverberation, unlike early reflections, is assumed to be consistent across a space. How correct this assumption is is dependent on the specific space including its geometry and absorptive materials [Blesser and Salter, 2007, p. 253]. This assumption implies ergodicity: consistent statistics across listening locations. When and if this noise-like, ergodic signal emerges, a room can be considered *mixed*.

The transition between the early reflections and late reverberation can be described in a number of ways and is often called the mixing time. Blesser [2001] states the mixing time is “how long it takes for there to be no memory of the initial state of the system. There is statistically equal energy in all regions of the space after the mixing time.” Here the mixing time is accepted to be after the transition from early reflections to late reverberation is complete and the earliest point in the IR that a stochastic rather than deterministic model can be used. The upper limit of mixing time has been discussed by Polack [1993] and further by Jot et al. [1997] to be proportional to the average distance between all pairs of points on a surface which is also proportional to the volume of a space divided by the surface area:

$$t_{mixing} \propto \frac{\sqrt{V}}{S} \tag{2.2}$$

where V is the volume and S is the surface area.

The transition or mixing of a space can also be described in terms of frequency, time, and distance. The Schroeder frequency is the frequency above which a single input frequency will excite at least three different resonances of the space [Schroeder, 1996]. This introduces more randomness into the system making rooms sound like reverberant spaces and not just comb filters. The Schroeder frequency is defined as:

$$f_c = 2000\sqrt{\frac{RT_{60}}{V}} \tag{2.3}$$

where RT_{60} is the reverberation time in seconds and V is the volume of the room in m^3. The Schroeder frequency can also be expressed as a wavelength:

$$\lambda_c = \sqrt{\frac{A}{6}} \tag{2.4}$$

where A is the equivalent absorption coefficient of the room. The absorption coefficient is a frequency-dependent measure of sound absorption. A table of select absorption coefficients for common materials found in building construction can be found in in [Beranek, 2004, p. 639–640].

The transition from a deterministic system to a stochastic one can be described in terms of the rate of echoes that arrive at a receiver. The diffuse-field time interval t_c is the longest time between echoes where the signal can be considered diffuse [Schroeder, 1996]:

$$t_c = \sqrt{\frac{ln(10^6)V}{4\pi c^3 RT_{60}}} \tag{2.5}$$

When examining the distribution of samples in an IR, the late reverberation tends towards a normal distribution, unlike the energy from early reflections. A progression towards a more normal distribution occurs as time increases and the acoustic energy within the space becomes more mixed [Blesser, 2001]. A measurement of distribution can then be used to determine whether a point in time is more or less deterministic, that is, whether it is within the early reflections or late reverberation.

Abel and Huang [2006] studied a measure called the echo density profile to observe the rate of reflections in a reverberant signal. The metric examines a sliding window of samples from an IR and is defined below in Equation 2.6. The echo density profile is the number of samples in the window that are outside the standard deviation of the distribution. The samples in the sliding window will tend towards a normal distribution as time progresses and the echo density increases. They also looked at kurtosis as a statistical measure instead of standard deviation, but felt that it often indicated false peaks when the window contained a reflection and did not perform well when the echo density is small.

$$\begin{aligned} \eta(t) &= \frac{1}{\operatorname{erfc}(1/\sqrt{2})} \sum_{\tau=t-\delta}^{t+\delta} \omega(\tau)\mathbf{1}\{|h(\tau)| > \sigma\} \\ \sigma &= \left[\sum_{\tau=t-\delta}^{t+\delta} \omega(\tau)h^2(\tau) \right]^{\frac{1}{2}} \end{aligned} \tag{2.6}$$

with $\mathbf{1}\{\cdot\}$ being the indicator function which returns one when the argument is true and zero when false. The complementary error function $\operatorname{erfc}(1/\sqrt{2}) \approx 0.3173$ is the expected fraction of samples outside one standard deviation and $\omega(t)$ is a windowing function, recommended by the authors to be a Hann window.

Hidaka et al. [2007] state "there should be a single physical measure for determining the point at which the early reflections have been overtaken by the late reverberant sound, i.e., the point at which the correlation between the two is as at a low value." To find this low value of correlation, the IR is split into two sections at different points in time. The point in time that creates two sections with the lowest correlation is considered the transition or mixing time. This measurement was applied to 57 halls which had a calculated transition time of 70 to 280 ms. However, no direct relation to the subjective sound of the hall was found.

2.2.5 Acoustic metrics

Whilst humans have a rich language to describe sound, particularly in regards to reverberation, reliably extracting descriptions of a reverberant environment automatically from a signal is not a solved problem. As can be seen in Table 2.2, various attributes of a room contribute to the perceived acoustical quality. A variety of metrics have been proposed which attempt to map signal to perception with the hope of improving room design. Beranek [2004] has compiled 25 common terms that are used by musicians, acousticians, and music critics to describe music performance spaces. Some of these terms can be easily described in mathematical terms; others are more abstract and difficult to

Perceptual Quality	*Physical Attribute*
Quality of Early Sound	Abundance of lateral reflections Low correlation between early sounds and the two ears at mid-frequencies
Reverberant Sound	Reverberation time optimum for type of music performed Adequate strength Irregular surfaces in space
Quality of Bass Sound	Adequate strength of sound in frequency range 80-355 Hz
Loudness	Space must not be too large Space must contain minimum of sound-absorbing materials
Clarity or definition	Proper ratio of direct sound energy to reverberant sound energy for the music being performed Restricted reverberation time Affected by speed of music
Timbre and tone colour	Affected by texture, balance, and blend of music in sound field Irregularities on surfaces in space Balance of tonal spectrum

Table 2.2: The dependency of aspects of musical performance in a space and the acoustical attributes of that space, adapted from [Beranek, 2004, p. 34] to include only terms relevant to this thesis.

measure scientifically. Though no single metric can fully describe a room, a combination of metrics can provide a starting point for automatic analysis. Here we will briefly review the metrics that are pertinent to this thesis.

Reverberation Time is perhaps the most common acoustic term, though its calculation is often misunderstood. The reverberation time is the time a sound takes to decay to 60 dB below the initial level once the sound source has stopped. Not every frequency will decay at the same rate and spaces complimentary to music generally take 1.8 to 2.0 s for frequencies between 350 and 1400 Hz to decay [Beranek, 2004, p. 21]. Frequency-dependent reverberation time is then a measure of the time-frequency envelope of reverberation.

Sabine [1992] was a pioneer of acoustics who derived the reverberation formula:

$$RT = 0.163\frac{V}{A}\text{seconds} \tag{2.7}$$

where V is the volume of the room in m^3 and A is the absorption coefficient.

When the reverberation time is derived directly from a measured IR, it is denoted as RT_N or $RT_{N,f}$ where N is a number of decibels and f is a frequency. Reverberation time is calculated from the energy-decay curve (EDC) of the IR, the backwards integration

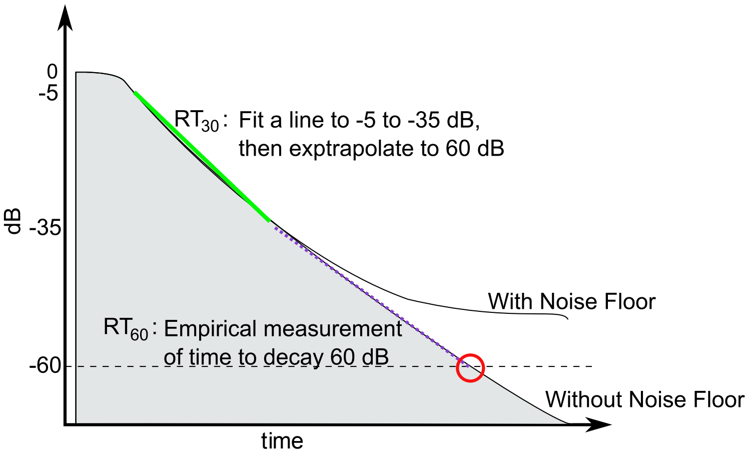


Figure 2.6: Illustration of how reverberation time is calculated from an energy decay curve.

of the IR as described below [Schroeder, 1965].

$$EDC(t) = \int_{t}^{\infty} h^2(\tau)d\tau \tag{2.8}$$

The $EDC(t)$ is the amount of energy remaining in the IR, h, at time t. RT_{60} refers to the measured time for the EDC to decay 60dB, but RT_{30} is not the time for the EDC to decay 30 dB. It still is a measure of time for the sound to decay 60 dB. RT_{30} is extrapolated from the line fit from -5 dB to -35 dB. This is illustrated in Figure 2.6. The same process of fitting a line and then extrapolating the time at 60 dB is used for any other value of RT_N where the line is fit to -5 dB through $-5-(N+5)$ dB. Values other than RT_{60} are computed in order to compensate for noise floors that are less than 60 dB below the peak of the measured IR.

Clarity measures the relative amount of early and late energy expressed in decibels with the distinction between early and late being 50 ms for speech (C_{50}) and 80 ms for music (C_{80}). Higher clarity values indicate a clearer sound or more understandable speech. [Beranek, 2004, p. 536] recommends C_{80} of -3.0 to 0.0 for symphonic repertoire and 1.0 to 3.0 for opera.

$$C_l = 10\log\left(\frac{\displaystyle\int_{0}^{l} h^2(t)dt}{\displaystyle\int_{l}^{\infty} h^2(t)dt}\right)\text{dB} \tag{2.9}$$

where l is the time limit of either 50 or 80 ms.

IACC is the interaural cross-correlation coefficient, a measure of the difference between the signals arriving at each ear. It is believed that the human brain keeps a running average of ITD cues through cross-correlation of the signals that arrive at each ear [Blauert, 1997, p. 168].

$$IACC_t = max|IACF_t(\tau)| \quad \text{for} - 1 < \tau < 1 \tag{2.10}$$

$$\text{where} \quad IACF_t(\tau) = \frac{\int_{t1}^{t2} h_L(t)h_r(t+\tau)dt}{\int_{t1}^{t2} h_L^2(t)dt \int_{t1}^{t2} h_R^2(t)dt} \tag{2.11}$$

The signals at the right and left ears are h_R and h_L respectively. Typically if the signal being analysed is an IR of a room, then $t_1 = 0$ ms and $t_2 = 80$ ms when looking at only early reflections and $t_1 = 80$ ms and $t_2 = \infty$ for the diffuse, reverberant sound. Otherwise $t_1 = 0$ ms and $t_2 = \infty$ [ISO 3382, 1997].

The $IACC$ of a naturally-occurring binaural signal ranges from 0 to 1. An $IACC$ value of 1 indicates a single auditory event will be heard; a value less than 1 may mean the signals will be interpreted as a single event but spatially wider. When the $IACC$ is around 0.4, the auditory events appears to be spread across the entire frontal plane. As the value approaches 0 (meaning that there is no coherence between the two signals) a listener will hear two separate sound sources [Blauert, 1997, p. 242].

2.3 Modelling Reverberation

The first audio recordings consisted of a single transducer that captured and stored the sound field without any control over further modifications. Solo musicians along with ensembles could do little more than manipulate their physical proximity and volume in an attempt to create the desired effect on the recording. However, musicians needed to crowd close to the transducer in order to overcome poor signal-to-noise ratios due to the recording medium along with attenuation from distance. This caused the recordings to have an unnatural sound as the effects of the room were lost [Blesser, 2001].

The introduction of electrical circuitry and stereo recording techniques revolutionised the ability to craft a new listening experience that not only included the environment of a concert hall; it could synthesize a virtual space that is not physically reproducible. This was accomplished with artificial spatial audio techniques, both monaural and multi-channel.

Artificial reverberation, reverberation that was not present during the initial capture of the audio recording, aims to give a sense of physical space to audio. Recordings without any reverberation are perceived as unnatural as sound sources and their subsequent reverberation are not normally separated in the natural world. In the majority of cases the space does not need to exactly replicate a physical space, but it can be implied. Early attempts at introducing artificial reverberation primarily included physical spaces [Goodfriend and Beaumont, 1959]. The dry, or non-reverberant, recording would be played over a loudspeaker in a room with a desirable reverberation, and a microphone placed elsewhere in the room would record the new reverberant sound. The rooms used for reverberation could be built for another purpose such as a concert hall or could be a dedicated room in a recording studio. These dedicated rooms, reverberation chambers, were expensive for recording studios as they required the same degree of acoustic isolation

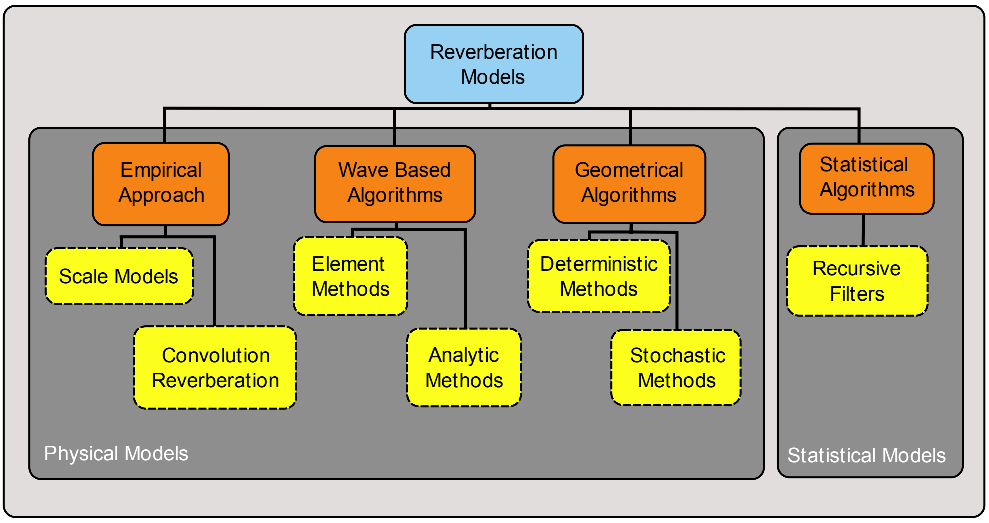


Figure 2.7: Taxonomy of sound propagation models, adapted from [Vorländer, 2008, p. 148]

as a room used for recording. Reverberation chambers also needed to have a pleasant sound, which as acousticians and architects will confirm, is neither a precise science nor simple task. Research efforts were then directed to artificial reverberation techniques with lower overhead costs and small physical footprints [Goodfriend and Beaumont, 1959, Moura and Campos, 1959].

While commercial music recordings usually only require an implication of a space, not an exact replication of a physical room, architects and investors of buildings such as concert halls do need an inexpensive, flexible, and precise model of a room. Scale models were the earliest method for simulating sound propagation in a space and are a technique still used [Blesser, 2001]. Models can range from ratios of 50:1 to 10:1 but require different compensations in air pressure and humidity. While building scale models is cheaper and involves less commitment than building a complete room, they still lack flexibility. Computer models can provide flexibility with even less cost.

Artificial reverberation can be created in a variety of ways with trade-offs of accuracy, flexibility, and computational speed. A taxonomy of classes of algorithms can be seen in Figure 2.7. In Sections 2.3.1 to 2.3.3 we will briefly cover the artificial reverberation techniques that are relevant to the subsequent chapters: convolution reverberation, statistical models, and physical models.

2.3.1 Convolution reverberation

An acoustic space can be assumed to be a linear, time-invariant (LTI) system, though it is acknowledged that real spaces do vary over time. When a LTI system is assumed, the system can be measured by an IR which will contain all needed information about the system. An IR sounds like a recording of a gunshot or balloon pop and the subsequent reverberation in the room being measured. When this recording is convolved with a dry sound, one without any reverberation, the illusion is created that the sound was recorded in that room. This method can create highly accurate measurements of a space and produce high-quality spatialisations. However, the virtual sound source and receiver positions are limited to the positions of the microphone and loudspeaker during the IR measurement.

Convolution of an IR with a dry sound source is described below. As can be seen in the equation, only a single IR is convolved meaning that only the information pertaining to that particular IR is contained in the output signal. Any change in the room size or construction materials or change in source or receiver position must be represented in a new IR.

$$y(t) = h(t) \otimes x(t) = \int_0^\infty h(\tau)x(t-\tau)d\tau \tag{2.12}$$

where the dry source signal $x(t)$ is convolved with the IR $h(t)$ and $y(t)$ is the reverberant signal.

The best practice for measuring an IR has evolved as researchers have developed more accurate methods to capture the IR of a room. Initially, starter pistol shots and balloon pops were recorded and used as IRs, but these methods do not produce reliable results as neither the frequency response nor directivity are ideal and the exact excitation signal cannot be exactly reproduced [Bradley, 1986]. The general technique most commonly employed is to excite the space with a known signal, record the result, and then remove the known signal. The known signal is usually a maximum-length-sequence (MLS) or a sine sweep. MLS is a pseudorandom binary sequence with a flat power spectrum [Vanderkooy, 1994]. Multiple measurements are usually taken and averaged in order to improve the signal-to-noise ratio. Non-linearities from the measurement system can be mistakenly measured if care is not taken [Holters et al., 2009].

Linear sine sweeps, also known as time delay spectrometry, are limited to short measurements and have a number of disadvantages which now overshadow its previous computational advantages. Farina [2000] developed a new log sine sweep which improves the signal-to-noise ratio and also removes harmonic distortion that could have been introduced by the measurement system. The equation to generate this signal is:

$$s(t) = \sin\left[\frac{\omega_1 \cdot T}{\ln\left(\frac{\omega_2}{\omega_1}\right)}\left(e^{\frac{t}{T}\ln\left(\frac{\omega_2}{\omega_1}\right)} - 1\right)\right] \tag{2.13}$$

where the resulting sine sweep, $s(t)$, is dependent on starting and ending frequencies ω_1 and ω_2 and has a duration of T seconds.

For an IR to be a comprehensive and absolute measurement of a room, that room needs to be linear and time-invariant; real rooms are seldom either. Temperature fluctuations are largely the issue as thermal turbulence causes unrepeatable variability in measurements [Blesser, 2001]. IRs also are often measured with directional sound sources with a flat frequency response or omnidirectional sound sources with a poorer frequency response, as it is currently impossible to have a sound source with omnidirectional directivity and a flat frequency responses. An omnidirectional, full bandwidth measurement is desired even though sound sources are directional. The

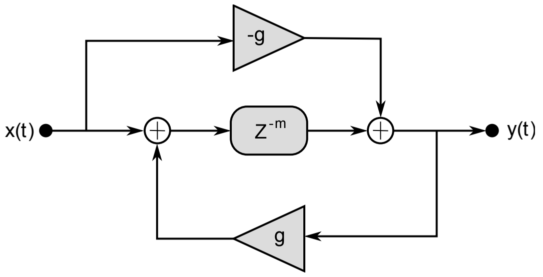


Figure 2.8: Schroeder all-pass filter.

directivity pattern of a virtual sound source can be applied to an omnidirectional measurement.

While these are valid issues to raise, they do not invalidate the use of IRs and convolution reverberation. IRs, while imperfect, are an essential tool for acoustics research as they provide a numerical representation of a room. Additionally, convolution reverberation may be forcing a time-variant system to be modelled as time-invariant, but it synthesises a perceptually realistic and relevant model of reverberation.

2.3.2 Statistical models

Statistical models of room reverberation are largely based on recursive delay lines and are intended to recreate the perceptual effects of late reverberation but not the physical reality of how sound interacts with a space. Statistical models aim to parameterise reverberation, often attempting to map signal processing parameters to perceptual attributes. In order to have complete control over the output of a statistical reverberator, it is necessary to have recursive delay lines which do not 'colour' or alter the frequency response of a sound.

Reverberation is a series of reflections that increase in density over time, theoretically approaching an infinite number of reflections. A logical model of such a system would be a simple delay line with a feedback loop. This, however, creates a comb filter which 'colours' the sound. Upon first inspection, comb filtering appears to be an accurate representation of reverberation as a room creates a comb filtering effect. However, the key difference is in the density of resonances. A single or series of delay lines have a constant density of resonances across the spectrum, but the concentration of resonances in a room increases with frequency. Delay lines alone alter the frequency content of the sound in an unrealistic and undesirable way.

Delay lines are necessary in order to increase echo density, but echo density needs to increase without altering the frequency response. In 1962, Schroeder found the solution: all-pass filters. All-pass filters mix delayed and undelayed signals in the correct ratio so that each frequency has unity gain at filter output; a Schroeder all-pass filter can be seen in Figure 2.8. While the frequency response is unaltered by an all-pass filter, the phase-response is a non-linear function of frequency resulting in smearing in the time domain [Gardner, 1998, p. 107]. A large range of reverberators can be designed by putting delay lines and all-pass filters in various configurations, but they still lack the

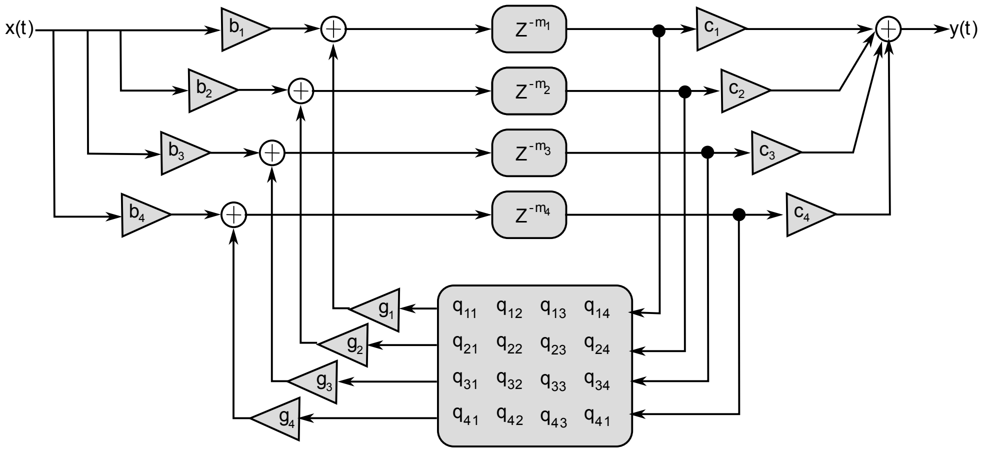


Figure 2.9: Diagram of a generic feedback delay network.

echo density needed for natural sounding reverberation. They also have a tendency to sound 'metallic', especially with transient signals [Moorer, 1979]. A linear system of delays and all-pass filters does not have a Schroeder frequency.

Gerzon [1976] published a mathematical approach to reverberation design that was later developed by Jot [1991]. The key aspect of this new reverberator is the unity-orthogonal mixing matrix which regulates the energy in the filter so that the energy leaving is always equal to the energy entering [Blesser and Salter, 2007, p. 264]. The reverberator structure is referred to as a feedback delay network (FDN) and can be seen in Figure 2.9.

As mentioned earlier, it is not possible to absolutely capture how sound moves within a space using an IR measurement as rooms are not truly linear time-invariant systems; randomness is introduced from air turbulence. Many commercial reverberators introduce time-variance as modulation of parameters like delay length to perceptually simulate real spaces [Blesser and Salter, 2007, p. 268].

2.3.3 Physical models

Physical models aim to precisely recreate a space. The accuracy of the model depends on the amount of detail recorded such as absorption coefficients, architectural details, and the mathematics calculating how the energy is propagated in the space. When a model is auralized, a very similar process to convolution reverberation occurs. Instead of convolving a sound source signal with a measured IR, the source is convolved with an IR rendered and output by the model.

Geometrical acoustics methods model sound as a ray or a particle, meaning wave effects such as diffraction and scattering are ignored [Vorländer, 2008, p. 58]. They assume that the sound source wavelengths are significantly smaller than the room boundaries, so they are not valid at low frequencies. These simplistic models were used extensively up to the middle of the 20^{th} century [Vorländer, 2008, p. 200]. Two basic geometrical models of a room are illustrated in Figure 2.10, image source and ray-tracing. Image source approaches work well for deterministic models while ray tracing can model

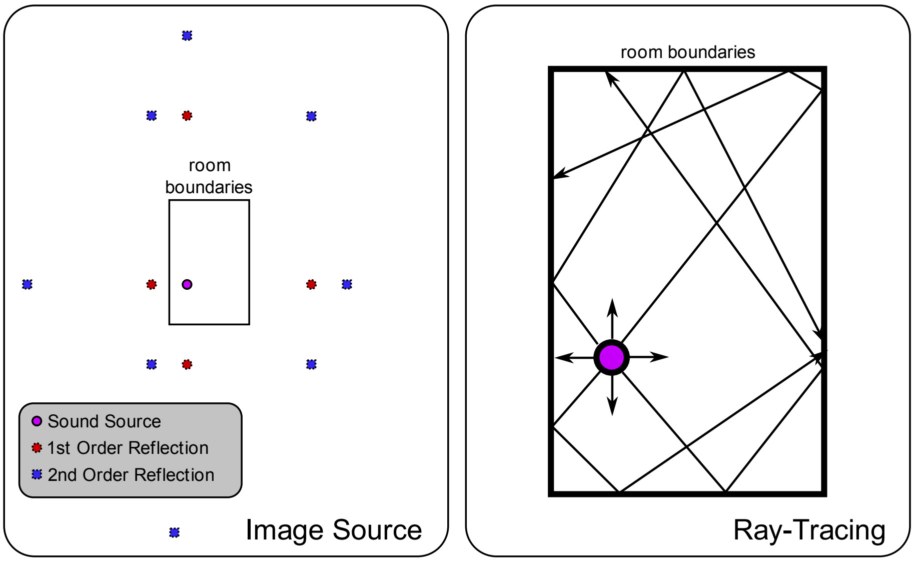


Figure 2.10: Image source and ray-tracing models of room reverberation.

stochastic portions of a room response; many commercial room modelling applications combine the two approaches into a hybrid model [Vorländer, 2008, p. 217].

Image source models are calculated by reflecting the original source across the room boundaries (first order reflections) and reflecting those reflections across the boundaries (higher order reflections). The number of reflections grow exponentially, so computation time usually limits the number of orders calculated [Vorländer, 2008, p. 204].

Ray tracing can be visualised as balls on a billiards table. A number of balls are sent in all directions from the sound source in the table which is shaped like the room being modelled. Each time a ball hits a side of the table, it is reflected to a new direction but it also loses some energy. When a ball reaches the virtual receiver in the room, the number of times the ball or ray hit a boundary is recorded [Vorländer, 2008, p. 181-182].

Wave based algorithms include finite element methods and waveguides. These models subdivide a space into smaller sections and then solve the wave equation for that subsection. Digital waveguide meshes (DWM) sample a space using a mesh structure and find a discrete approximation to the first order solution of the wave equation. Finite element methods divide a volume into a series of smaller subvolumes and use a discrete approximation to the wave equation [Vorländer, 2008, p. 156].

2.4 Loudspeaker Rendering

As discussed above, there is a variety of approaches to synthesise a space. There are also multiple techniques to convey that space to a listener. This section presents a sampling of those techniques that use loudspeakers. Loudspeakers have the advantage of conveying the same sound field to multiple people. There are limits to size and position of the ideal listening area where the spatial sound field is best reconstructed. This area is called the sweet spot and its attributes vary according to the reproduction system. The sweet spot is often limited by the precedence effect which can cause sounds to appear to be

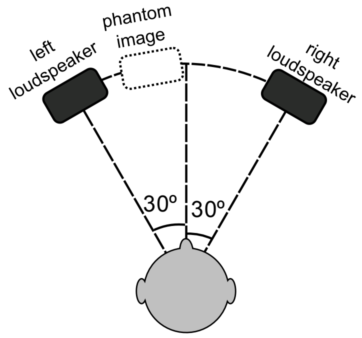


Figure 2.11: Two-channel panning.

emanating from the closest speaker, regardless of its intended virtual position.

2.4.1 Panning

Panning manipulates the phase or amplitude relationships amongst multiple loudspeaker outputs to mimic ILD and ITD cues. Amplitude panning, which takes advantage of ILD cues, is by far the most popular method of synthesising a spatial sound field and is often simply referred to as 'stereo.'

An amplitude difference between 15 and 19 dB at frequencies below approximately 700 Hz moves a sound completely into the loudest speaker of a stereo pair, but subtler differences between two loudspeakers can create a 'phantom image' [Malham, 1998]. Equation 2.14 describes the relationship between the left and right channels that feed two loudspeakers and the location of the desired phantom stereo image between the speakers.

$$\frac{\sin\theta}{\sin\theta_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{2.14}$$

where g_1 and g_2 are the loudspeaker gains, θ is the angle between the listening position and each loudspeaker, and $\pm\theta_0$ is the resulting phantom image angle between the median plane and the loudspeakers. These relationships are further illustrated in Figure 2.11.

Stereo panning has been extended to larger loudspeaker systems and is used in 5.1 systems. Multiple loudspeakers systems such as 5.1 or 10.2 use a common notation. The figure preceding the decimal indicates the number of full bandwidth loudspeakers and the figure following the decimal denotes the number of low-frequency effects channels. Large loudspeaker systems such as the 5.1 system that have grown out of the film industry often use panning to place sound sources around the virtual space, but since these systems do not have a uniform layout of loudspeakers, this is not completely effective [Rumsey, 2001, p. 86-89]. Cinema-based systems place more loudspeakers directly in front of the listening audience and spread fewer loudspeakers farther apart to the sides and rear. As the distance and angle between loudspeakers increases, panning provides a less robust illusion [Blesser and Salter, 2007, p. 210]. While this may be acceptable for enhancing the viewing of a film, it usually is not desired for musical listening.

Coincident microphone recording techniques employ ILD cues to create a stereo

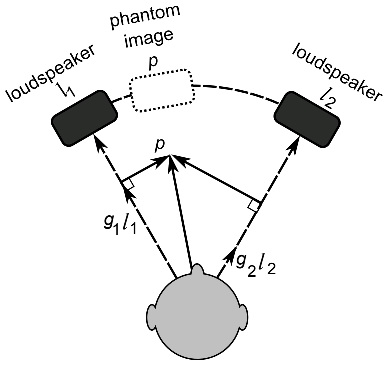


Figure 2.12: Illustration of VBAP.

image with two channels. The microphones' diaphragms are theoretically at the same point in space, so only differences in amplitude, not time, exist [Huber and Runstein, 2001, p. 118-119]. While time or phase differences are less exploited in synthetically generated audio, they are employed in stereo recording with spaced omni-directional microphone recording techniques [Malham, 1998].

2.4.2 Vector-base amplitude panning

Vector-Base Amplitude Panning (VBAP) creates a two or three-dimensional sound field. It aims to create a sound field with an unlimited number of loudspeakers placed arbitrarily around a listener. The only constraints are that the loudspeakers remain equidistant from the listener and that the room containing the loudspeaker system is relatively dry without distracting reflections. We are only giving a brief summary of VBAP here, but more detail can be found in [Pulkki, 1997, 2001a].

With the simplest case of two loudspeakers, VBAP is the same as amplitude panning, but the position of the phantom image is expressed as vectors as seen in Figure 2.12. VBAP is defined by unit-length vectors pointing towards loudspeakers 1 and 2, $l_1 = \begin{bmatrix} l_{11} & l_{12} \end{bmatrix}^T$ and $l_2 = \begin{bmatrix} l_{21} & l_{22} \end{bmatrix}^T$; T is the matrix transposition. A linear combination of the loudspeaker vectors creates vector $p = \begin{bmatrix} p_1 & p_2 \end{bmatrix}^T$ which points towards the phantom image.

When VBAP is implemented on a two-dimensional sound system with more than two loudspeakers, only two loudspeakers are used at a single time to generate a single source. The loudspeakers are paired and each loudspeaker can belong to two pairs. The vector p is then calculated as described above using the two closest loudspeakers to the virtual position.

By adding loudspeakers above or below the two-dimensional setup, three-dimensional VBAP can be used. The loudspeakers need to form triangular configurations and need to be equidistant from the centre listening position. A virtual sound source is then panned amongst a triangle of loudspeakers instead of a pair of loudspeakers. The unit-length vector is extended to $l_1 = \begin{bmatrix} l_{11} & l_{12} & l_{13} \end{bmatrix}^T$. The three-dimensional vector p is a linear combination of vectors l_1, l_2, and l_3. The gain factors for each loudspeaker for a given vector p are expressed in Equation 2.15.

$$\begin{bmatrix} g_1 & g_2 & g_3 \end{bmatrix} = p^T L_{123}^{-1} = \begin{bmatrix} p_1 & p_2 & p_3 \end{bmatrix} \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix}^{-1} \tag{2.15}$$

2.4.3 Ambisonics

Ambisonics can be considered both a microphone and a synthesis technique. It is based on spherical harmonics and directional information that can be encoded into as few as four channels to represent three-dimensions Fellgett [1975], Gerzon [1973, 1975]. Higher-order Ambisonics uses more channels to represent a directional sound field [Rumsey, 2001, p. 111-117]. Table 2.3 lists the channel name, an illustration of the spherical harmonic, and the encoding equation for each B-format channel.

As a microphone technique, the first order spatial information of a recorded signal is encoded into four signals: an omni-directional signal W, a figure of eight signal X capturing front/back information, a figure of eight signal Y capturing left/right information, and a figure of eight signal Z capturing up/down information. These four channels are referred to as B-format [Malham, 1998]. First order B-format signals can also be synthesised using the coefficients in Table 2.3.

The coefficient 0.707, approximately $1/\sqrt{2}$ in the calculation of W, is an engineering adjustment made to balance the distribution of the levels for each channel. The B-format signal does not include any distance cues which Malham [1998] suggests including by controlling loudness and direct-to-reverberant ratios of sound sources.

Higher order Ambisonics are extensions of the same principles as first order, but use higher order spherical harmonics which allow for greater localisation accuracy in recording and reproduction. The resulting B-format encoding contains more channels as the order increases. Table 2.4 lists the channels required for first through third order encoding for 2D and 3D sound fields and Table 2.3 lists their encoding coefficients.

If all channels in the B-format signal are processed equally, then the directional information is preserved amongst the channels [Malham, 1998]. The sound field can be transformed, e.g. rotated about an axis, without disrupting the original directional relationships between sources. An example of a transformation matrix is below. The matrix rotates the entire first order sound field around the Z axis by an angle β.

$$\begin{pmatrix} \cos\beta & -\sin\beta & 0 \\ \sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.16}$$

An Ambisonics decoder produces discrete signals to be fed to individual loudspeakers to recreate the directional sound field. Ambisonics is advantageous over standards such as 5.1 surround sound as there is a multitude of possible loudspeaker arrangements and methods to compensate for less-than-ideal configurations. The minimum number of speakers required for decoding is four in a two-dimensional rectangle and eight in a three-dimensional cube. In general, the greater the number of speakers, the better the

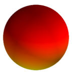
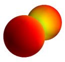
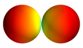
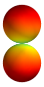
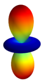
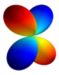
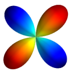
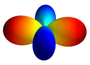
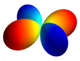
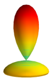
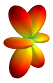
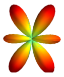
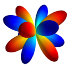
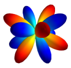
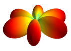
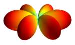
Order	*Channel*	*Spherical Harmonic*	*FuMa Coefficient*
1st	W		$1/\sqrt{2}$
	X		$\cos(\theta)\cos(\phi)$
	Y		$\sin(\theta)\cos(\phi)$
	Z		$\sin(\phi)$
2nd	R		$3\sin^2(\phi-1)/2$
	S		$\cos(\theta)\sin(2\phi)$
	T		$\sin(\theta)\sin(2\phi)$
	U		$\cos(2\theta)\cos^2(\phi)$
	V		$\sin(2\theta)\cos^2(\phi)$
3rd	K		$(1/2)\sin(\phi)(5\sin^2(\phi)-3)$
	L		$\sqrt{135/256}\cos(\theta)\cos(\phi)(5\sin^2(\phi)-1$
	M		$\sqrt{135/256}\sin(\theta)\cos(\phi)(5\sin^2(\phi)-1)$
	N		$\sqrt{27/4}\cos(2\theta)\sin(\phi)\cos^2(\phi)$
	O		$\sqrt{27/4}\sin(2\theta)\sin(\phi)\cos^2(\phi)$
	P		$\cos(3\theta)\cos^3(\phi)$
	Q		$\sin(3\theta)\cos^3(\phi)$

Table 2.3: Furse-Malham coefficients for Ambisonics, see Malham [2009] for discussion of the coefficients. θ = source azimuth, ϕ = source elevation

Order	*Sound Field Type*	*Number of Channels*	*Channel Names*
1st	horizontal	3	WXY
	spherical	4	WXYZ
2nd	horizontal	5	WXYUV
	spherical	9	WXYZRSTUV
3rd	horizontal	7	WXYUVPQ
	spherical	16	WXYZRSTUVKLMNOPQ

Table 2.4: Channel names associated with first through third order B-format.

reproduction as long as they are evenly distributed around the central listening position [Malham, 1998].

Ambisonics decoders increase in complexity when trying to compensate for factors such as large listening areas. At its simplest, a decoder distributes the B-format signal to each loudspeaker according to its position from the centre of the listening area using a decoding matrix [Malham and Myatt, 1995]. The loudspeaker feeds are the signals generated by a virtual microphone placed where the loudspeaker would be. Below is an example of how a loudspeaker feed is computed from the azimuth θ and elevation ϕ and a first order B-format signal.

$$L_{\theta,\phi} = 1/\sqrt{2} \times W + \cos\theta\cos\phi \times X + \sin\theta\cos\phi \times Y + \sin\phi \times Z \tag{2.17}$$

Decoders ideally should also include dual-band decoding to optimise localisation using the velocity vector r_v and the energy vector r_e as discussed in [Gerzon, 1992]. If the loudspeakers are within a few meters of the listener, the decoder should also compensate for near-field effects [Daniel, 2003]. A lengthy discussion of designing Ambisonics decoders for smaller spaces can be found in [Heller et al., 2008], and a discussion of decoders for larger spaces can be found in Malham [1992].

2.4.4 Wave field synthesis

Wave field synthesis (WFS) attempts to accurately reconstruct the wave field of a sound source. It is based on Huygen's Principle which states that any complex wave field can be reconstructed with basic waves [Berkhout et al., 1993]. As illustrated in Figure 2.13, when the wave field is recorded by microphones at different spatial positions, and each of those recordings are played back from loudspeakers spatially distributed as the microphones were, the sound field will be reconstructed.

If spatial sampling is being considered, a perfect reconstruction of frequency content up to 10 kHz in a 30 m^3 space would require around 50,000 loudspeakers, each about 3 cm in size. This is impractical for a number of reasons, not least being cost. When psychoacoustical approximations are considered along with other simplifications, typical two-dimensional arrays using around 500 loudspeakers are sufficient [Vorländer, 2008, p. 290-291].

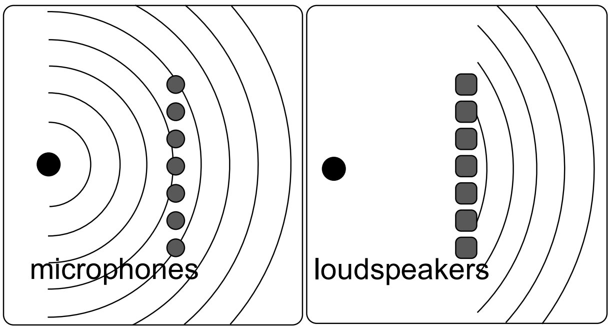


Figure 2.13: A simplification of the underlying concept for WFS.

2.5 Binaural Headphone Rendering

Binaural audio is spatial audio that is intended for headphone playback. In general, binaural audio should not be played over loudspeakers as the effects of head shadowing and reflections off the pinnae are encoded in the audio. This includes the ITD, ILD, and frequency cues discussed in Section 2.1. Any musical performance or speech can be encoded into binaural audio by recording with two microphones inside either a dummy head or from within a listener's ears. The simplest dummy head is a sphere with a microphone on either side to approximate a head, but models can include intricate pinnae and a full torso. Binaural recordings usually do not produce a convincing spatial audio image. The sound field does not move with the listener's head and as was discussed in Section 2.1.6, head movements provide important cues. Synthesis of binaural audio as opposed to directly recording an audio event is more flexible and head movements can be incorporated.

To synthesise a binaural sound field, virtual sound sources need to be convolved with head-related transfer functions (HRTFs). HRTFs are measured in a similar way to IRs described in Section 2.3.1. Microphones are placed in the ears of a human or a dummy head and a sound source is recorded from discrete locations on the median, frontal, and horizontal planes. They are typically recorded in an anechoic room in order to avoid room reflections. HRTFs encode ITD, ILD, and frequency cues and theoretically contain all of the information needed to simulate a sound emanating from a particular location in space. HRTFs have been found to be minimum-phase, so minimum-phase filters can be used to approximate measured HRTFs [Cheng and Wakefield, 2001].

Binaural audio can use a listener's own set of HRTFs or a set which was not measured with the listener's ears called non-individualised HRTFs. Synthesised binaural audio introduces a number of issues, the most common being front-back confusion: believing a sound source to be located in front when it is located behind or vice versa. Other errors include azimuth errors, elevation errors, problems convincing the listener that the source is located outside of their head (externalisation). Though these errors can occur when using a listener's own HRTFs, they can be further exacerbated with non-individualised HRTFs, but this is not consistent. When using speech as a stimulus, non-individualised and individualised HRTFs perform similarly [Begault et al., 2001]. When listeners are asked to locate a broadband noise source,non-individualised HRTFs can create up to four times more confusions between whether a sound is in front or behind the listener than with individualised HRTFs [Wenzel et al., 1993].

HRTFs are not typically recorded with room information; HRTFs combined with room IR measurements are referred to as binaural room IRs (BRIRs). Including some type of reverberation has been shown to improve azimuth and elevation errors and greatly aids in externalisation. There is not a significant difference in using only early reflections versus full reverberation with diffuse late reverberation, so a minimal acoustic field can be used [Begault et al., 2001].

2.5.1 Head tracking

Head movements can reduce front-back reversals if the sound field remains stable [Begault et al., 2001]; this is done with head tracking. The angle of rotation of the listener's head is tracked by a computer, and the binaural audio sound field is adjusted so that the sound sources appear to remain at the same location. For example, if a sound source is at an azimuth of 30° to the right of the listener and the listener turns 50° to the left, the sound source will then be placed at 80° to the right by the spatial audio algorithm.

Cone of confusion and median plane errors increase with non-individualised HRTFs, usually experienced as the sound collapsing inside the head [Moore, 2003, p. 251], [Cheng and Wakefield, 2001]. Compensating for head movements may help externalise sound sources. Head movements help resolve sounds on the cone of confusion or median plane, but if other cues place the sound inside the head, then head-tracking alone will not externalise the sound [Begault, 2000, p. 80]. Some head movements are unconscious and minute. Wersényi [2009] found that randomly moving a sound field 1° to 2° can significantly improve the externalisation of sounds sources, though it has no effect on front-back reversals.

2.5.2 Interpolated HRTFs

A set of HRTFs represents discrete points in space at different azimuths and elevations around a listener's head. A virtual sound source, however, may need to be placed at a location where an HRTF was not measured. This is especially common if head tracking is used. The measurement locations of HRTFs quantise the virtual space around a listener. A sound source location can be approximated by the closest HRTF if it is believed that the listener would not be able to tell the difference between the intended location and the rendered one.

In some cases, new HRTFs are estimated between measured HRTF points using interpolation. All interpolation algorithms introduce localisation errors, but it has been found that linear interpolation is sufficient when large, equidistant measurements are present; only small improvements are found with other interpolation methods [Nishino et al., 1999]. See Begault [2000, p. 132-136] for a more detailed discussion. The most significant factor for decreasing error is computing the interpolation in the frequency rather than the time domain [Hartung et al., 1999]. Figure 2.14 illustrates the general signal flow for synthesising a binaural sound field with interpolated HRTFs.

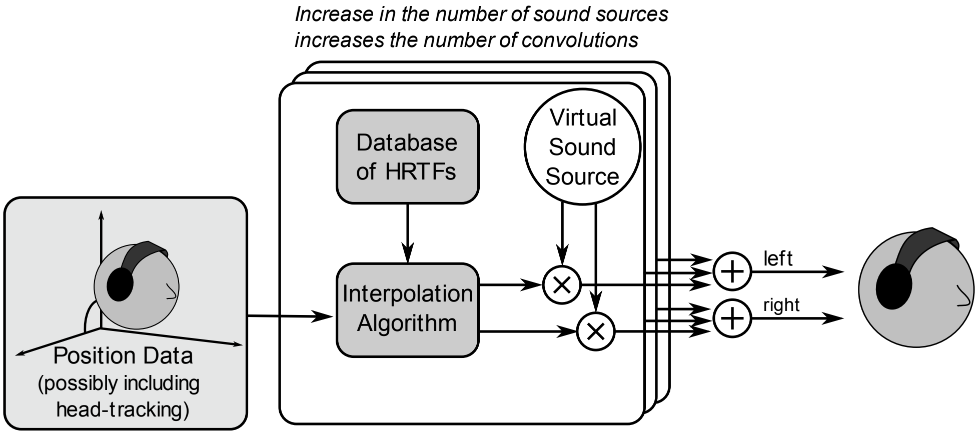

Increase in the number of sound sources increases the number of convolutions
Database of HRTFs
Virtual Sound Source
Interpolation Algorithm
Position Data (possibly including head-tracking)
left
right

Figure 2.14: Generating binaural audio by interpolating HRTFs. This signal flow is repeated for every individual virtual source.

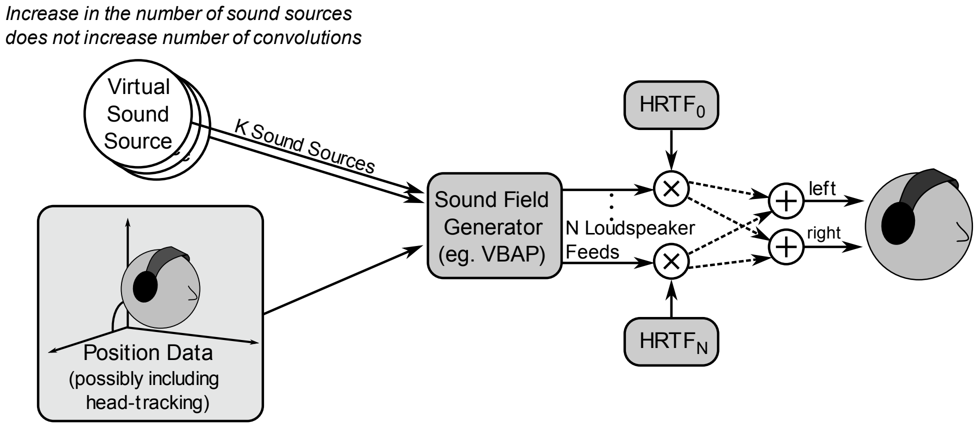

Increase in the number of sound sources does not increase number of convolutions
Virtual Sound Source
K Sound Sources
Sound Field Generator (eg. VBAP)
N Loudspeaker Feeds
HRTF$_0$
HRTF$_N$
left
right
Position Data (possibly including head-tracking)

Figure 2.15: Generating binaural audio with virtual loudspeakers.

2.5.3 Virtual loudspeakers

Real-time convolution of multiple audio sources no longer faces the computational constraints present in the 1990s. Modern desktop computers can easily accomplish what dedicated hardware was once required to do. Still, large reserves of memory and clock cycles are again important when considering mobile computing. Additionally, even though 3D graphics in video games are becoming more sophisticated in gaming consoles, audio processes are only given marginal memory and processor allotments.

The issues surrounding interpolating an HRTF set can be minimised when using a virtual loudspeaker approach. Instead of convolving each individual sound source with its appropriate HRTF in order to place it in the virtual sound field, an intermediary algorithm is used. The algorithm could be any discussed in Section 2.4 such as panning or Ambisonics. The loudspeaker signals calculated by the intermediary algorithm that would be fed to loudspeakers are instead convolved with the HRTFs associated with the desired loudspeaker position.

Ambisonics is extended to virtual Ambisonics when the decoded loudspeaker signals are convolved with HRTFs instead of fed to loudspeakers. The primary computational advantage of this approach is that a static set of HRTFs are used; there is no need for interpolation [Noisternig et al., 2003b]. The sound field is encoded into Ambisonics and all operations such as rotations for head-tracking are done within the B-format domain. Virtual Ambisonics also solves some practical limitations encountered by Ambisonics such as physical loudspeaker placement and listening position restraints.

Virtual loudspeakers have recently gained popularity in the research community, building from work in the 1990s which primarily relied upon the Lake DSP Convolvotron. McGrath and Reilly [1996] published one of the first systems using virtual Ambisonics on dedicated hardware. Virtual loudspeaker systems have also been used with a number of intermediary algorithms such as VBAP [Laitinen, 2008] and panning in commercially produced 5.1 recordings [Scott and Roginska, 2008].

2.6 Summary

Controlled laboratory listening tests have determined that a fundamental set of auditory cues inform our understanding of human perception of sound. These cues are largely built on the differences in timing, intensity, and frequency between the signals arriving at the two ears. Higher-level cognition plays a large role, but Duplex Theory is the basis for much of sound localisation. Additional or alternative cues are necessary as ITD and ILD cannot disambiguate every potential location such as those on the cone of confusion and the median plane. Frequency cues provide further information, but head movements allow for a more robust localisation as cues can be disambiguated by altering the relative location of a sound source.

By using these lower level cues in conjunction with higher cognition processes, humans are able to evaluate their surrounding spatial audio environment by ignoring information that appears to be redundant or extraneous and focusing on the pertinent sound source or sources. This includes subconscious suppression of multiple echoes from the same sound source created by environmental reflections and conscious suppression of interfering sound sources within a noisy environment. This means multiple sound sources can be presented to a listener within a reflective room and the listener will be able to distinguish between the sources and their reflections. A listener can even pay attention to a single source without interference from the room and competing sources.

These cues are limited and humans have a restricted spatial resolution for localising sounds. Sounds directly in front of the listener are located the most accurately, but sounds that move towards the sides or rear of the listener are more difficult to precisely locate. The precision varies from person to person and is dependent on the signal characteristics of the sound source and the larger context of the space.

Basic auditory cues can be manipulated to synthesise a spatial sound field primarily by varying phase and amplitude information. Loudspeaker-based spatial audio systems combine delayed and attenuated or amplified signals to simulate the sound field for a particular listening area. The listening area varies in size according to the loudspeaker system and encompassing room, but generally such systems are intended for multiple listeners. This is in contrast to binaural listening systems which are intended for headphone listening. As will be seen in Chapter 6, headphone listening does not need to be isolating and can be part of a shared auditory experience, but the social aspects of listening need to be considered when designing a spatial audio environment. In general, if the audio environment needs to be low-cost and/or portable then headphones are the preferred playback method. Either stereo panning or binaural audio can be used. If the audio environment needs to be shared simultaneously by multiple people, then loudspeakers should be used. The desired size of the sweetspot, the number of available loudspeakers, and limitations on where they can be placed around a physical space will determine the best rendering approach.

In addition to localising individual sources on the median, frontal, and horizontal planes, humans can evaluate the distance of individual sound sources and the surrounding space. This information is largely derived from reverberation, but with the exception of WFS, spatial audio rendering systems do not inherently consider distance. Most spatialisation algorithms strive to accurately place a sound at a particular angle from the listener but not at a specific distance. Additional cues need to be synthesised in order to simulate distance, the predominant cue being reverberation from an artificial reverberator. When synthesising a spatial sound field, reverberation is a vital factor in creating a realistic and immersive environment. The purpose of the virtual environment dictates what type of artificial reverberation is appropriate. Statistical reverberators are best for late reverberation and when a general aesthetic of space is needed, but a specific room does not need to be implied. For some applications, a precise model of a room is required, though if the model is required to render sound in real-time, the model is usually limited to geometrical models. Chapter 4 examines an application which requires a specific room to be modelled and develops a new signal processing approach that uses a more complex model in real-time.

Whilst reverberation creates a more immersive environment and decreases inside-the-head errors for headphone listening, it diminishes accuracy in localising a sound source. Despite this, Shinn-Cunningham [2000] has found that reverberation can be learned by listeners and some localisation accuracy returns with prolonged exposure to a virtual environment. Spatial audio cognition is dependent on learning and familiarity, but the learning process can be quick: within seconds upon entering an unfamiliar environment [Moore, 2003, p. 252]. Studies have shown that if subjects are not allowed to become familiar with a sound source or room their ability to localise a source decreases. The most common errors are similar to those that occur with non-individualised HRTFs: inability to localise sounds on the median plane and lateralisation inside the head instead of externalisation.

Unfamiliar environments also need to align with listener experiences in previous environments. Echoes are only suppressed by the precedence effect when they are consistent with listeners' expectations [Moore, 2003, p. 256]. If an echo is too different from the first occurring sound, then that sound will no longer be perceived as a repeat of the first event but an independent one. Learning and familiarity are important when synthesising a new virtual audio environment. The audio presented to a listener needs to fulfill expectations that have been learned from the physical world, but new events can be expected and learned if they are consistently presented. Chapter 3 establishes whether a virtual loudspeaker approach as discussed in Section 2.5.3 accurately recreates the spatial cues captured by HRTFs. If these cues are indeed reproduced, then a virtual loudspeaker approach using Ambisonics is an efficient and effective means to render a virtual audio environment.

A spatial audio environment with multiple complex streams of audio can be understood by a listener under certain conditions. If these conditions, such as spatially separating sound sources and using a perceptually valid distance model, are met then the environment can be used to convey information to a listener. If listeners are allowed to become familiar with the environment, they can learn to better interact and localise sound sources within that environment. Multiple streams of audio in a virtual spatial audio environment are used to convey complex information about a collection of music in Chapter 6. When listeners become accustomed to the environment, they navigate the virtual space to discover new music.

A spatial audio environment can be simulated with an artificial reverberator. The reverberation algorithm can be convolution with an IR of an existing space or be a model of a specific or generic space.

Chapter 3

Optimising Binaural Auditory Display

Visualisation of data is perhaps the most common means of exploring a collection of information. Charts and graphs give a sense of space and relationships between objects or data. However, not all information is suited for visual presentation – information may be better presented aurally. Particular types of information are easier to understand when they are heard rather than seen. This includes room acoustics models and music. Looking at a graph plotting an IR of a room may help identify or examine a particular aspect of the data, but vital information in the signal may be most obvious when heard. The same is true with music. An elaborate graphic of how different pieces of music are related may be visually interesting, but the information may be best absorbed and evaluated when listening to those songs.

Auditory display is the presentation of information with audio and is similar to visual display, which is the presentation of visual information. Auditory display can be used to present types of information that initially may not be associated with audio. This includes system monitoring and warning messages. Auditory display can be applied to a large range of tasks; it can reinforce information presented via another mode, typically visually, or can introduce new information. This makes it complementary to mobile computing where visual information is limited or seldom available [Bre, 2001].

Spatial auditory display is auditory display which uses spatial audio to convey information. It is also referred to as virtual acoustic display and involves the auralization of data. The previous chapter outlined the basics of human hearing and how auditory cues can be manipulated to synthesise a virtual sound source at an arbitrary point in space. This chapter discusses how a virtual space can be used to create a spatial auditory display. Sections 3.1 and 3.2 consider the general signal processing requirements for spatial auditory display and determine that virtual Ambisonics can be an efficient and flexible means to synthesise a virtual environment. Section 3.4 goes on to evaluate whether virtual Ambisonics can effectively create a binaural auditory display. Auditory display is only introduced here, but encompasses a wide array of topics. Kramer [1994] is recommended for a thorough discussion of the field.

3.1 Auditory Display Design

Spatial auditory displays are used in a variety of applications. Medical applications include sonification of brain electrical activity for monitoring and diagnosis [Jovanov et al., 1999], experimental study of sound localisation, and clinical spatial hearing tests [Shinn-Cunningham, 1998]. The American National Aeronautics and Space Administration (NASA) has lead much of the research regarding spatial auditory display as spatial auditory display has a number of applications for operation and navigation of aircraft [Wenzel, 1992, 1996, Begault, 2000]. NASA's initial spatial auditory display research started around the same time the Massachusetts Institute of Technology (MIT) also began work in the field. While the NASA programme concentrated on accurate presentation of information, the work at MIT focussed on efficient presentation of multiple speech signals [Arons et al., 1989] and the navigation of audio recordings and broadcasts [Arons, 1993, 1997, Schmandt and Mullins, 1995, Kobayashi and Schmandt, 1997]. Much of the current spatial auditory research focuses on mobile communication and interaction with devices with small or no screens [Walker and Brewster, 1999, Borß and Martin, 2008]

Spatial auditory display may be implemented using any of the spatialisation methods discussed in Section 2.4 and 2.5 which discuss loudspeaker and headphone rendering of spatial audio. The requirements of the task to be performed determines the most appropriate rendering system. In general, systems which require multiple people to interact in the same physical space should use a loudspeaker approach while systems which require mobility or less equipment should use headphones.

While particular tasks require different approaches to the design of an auditory display, Frauenberger and Stockman [2009] have compiled some common design features often considered:

Mapping – use of data or information to change properties of a sound, e.g. mapping stock values onto pitch of a particular sound.

Events – non-speech sound events, covering the range from alarms to more complex sounds.

Continuous sound – any sound that is not a sound event, but used over a longer period of time in the interface.

Background – sound that is intentionally designed to go in the background, i.e. not attracting the highest level of attention.

Parallel – the use of multiple sounds simultaneously.

Themes – the use of sound families that have a similarity making them part of a functional group of sounds.

Semantics – sounds that are chosen for their semantic relationship with the information they represent.

Spatial auditory display can require additional considerations as there is often spatial movement in the virtual space by the user or other elements. Cohen [1991] identifies three interaction paradigms that can be adopted by a spatial auditory display designer.

Museum metaphor/orientation perspective – sources are stationary and listener is moving.

Theatre metaphor/egocentric perspective – sources are moving and listener is stationary.

Cocktail party metaphor/dancing perspective – sources and listener are moving.

While these are not exhaustive lists of what an auditory display may entail, they are representative of the types of design approaches considered. Further considerations include what kinds of information are being conveyed and within what kind of virtual environment. For a further discussion of different auditory display applications and design techniques, see [Barrass and Frauenberger, 2009].

3.1.1 Direct and indirect sonification

The information conveyed in an auditory display can be classified as direct or indirect sonification of data. Fernström [2005] defines direct sonification as when "the information sought is also the information displayed." Direct sonification occurs when the data being sonified is an audio signal or any other signal which can be played at an audio sampling rate. In such cases the data does not require an intermediary mapping to a set of audio parameters. Indirect sonification then occurs when the information sought is not the information displayed, but instead an abstraction of the underlying information. Indirect sonification requires varying levels of abstraction to represent data and actions. Historically, computer science has been concerned with indirect sonification such as augmenting or replacing visual displays with auditory information within human-computer interaction. Electrical engineering has traditionally examined auditory display with direct sonification as the verification of signal processing algorithms like acoustic models (often called auralization [Kleiner et al., 1993]). However, both fields can benefit from the other and a flexible signal processing platform should be able to support both.

Auditory icons and earcons are two non-speech approaches to indirect sonification. The term earcons broadly refers to indirectly sonified sounds. Abstract earcons are sounds not associated with a physical action and could be one or a series of musical pitches such as the sound accompanied with an error message in most operating systems. Blattner et al. [1989] say "[a]rt is to icons as music is to earcons. . . Even though earcons are not music, a knowledge of music is important in understanding the problems of designing audio messages." Hermann and Hunt [2005] point out that "[t]he simplest auditory display conceptually is the auditory event marker, a sound that's played to signal something (akin to a telephone ring). Researchers have developed the techniques of auditory icons and earcons for this purpose, yet they're rarely used to display larger of complete data sets."

Auditory icons are a subset of earcons and are referred to as representational earcons [Blattner et al., 1989]. Auditory icons are "caricatures of naturally occurring sounds" that are intended to bear some direct relation to what they represent [Gaver, 1986, 1989]. A well known example is the sound of crumpling paper when emptying the trashcan in the Mac OS. Auditory icons range from *symbolic*, where arbitrary mappings that rely on social convention, to *nomic*, where the meaning depends directly on the physics of the situation, to *metaphorical*, where there are similarities between the represented object and the representing sound [Gaver, 1986].

3.1.2 Environmental realism

When spatial audio is used in an auditory display, it may be used to create a realistic virtual environment or it may create a virtual space that would never exist in the real world. A realistic virtual space requires perceptually accurate rendering of sound source distances, azimuths, and elevations along with a full room acoustics model. An unrealistic virtual space only needs to imply a space and often exaggerates realistic auditory cues.

> A virtual acoustic display can be defined as a medium for accurately transferring information to a human operator using the audio modality; it combines directional and semantic characteristics to form naturalistic representations of dynamic objects and events in remotely-sensed or simulated environments. As with visual displays, this definition does not necessarily mean that the virtual representation must be indistinguishable from reality. Rather, it implies that the display should provide a functional equivalence to human audition in the context of the task to be performed. [Wenzel, 1992, p. 261]

The "task to be performed" should drive the perceptual requirements of the system, and an ideal platform creating spatial auditory displays should be able to accommodate a variety of tasks. This means it should be able to render complex acoustic scenes complete with early reflections and late reverberation that describe a specific space and also be able to create simpler spatial environments with or without generic reverberation.

The DIVA system [Huopaniemi et al., 1996] is an example of a realistic auditory display integrated into a virtual reality world . Room acoustics are dynamically rendered so that a listener can explore a virtual musical performance. Early reflections and late reverberation are rendered to replicate a specific acoustic space. Walker and Brewster [1999] describe a very different spatial auditory display which creates a virtual space for navigation of menu items and other common tasks required in mobile phones. The display does not try to recreate a specific room, but only indicate where sound sources are located. While binaural audio is used in both applications, the first application creates an accurate rendering of an acoustic space while the second does not create a realistic virtual space but places earcons on a virtual sphere surrounding the user. Both are considered spatial auditory displays and they convey information via audio, but the reverberation is included in the pertinent information being conveyed in the first system while it not an essential piece of information in the second.

3.2 Spatial Auditory Display Requirements

To summarise the above points, a spatial auditory display needs to accommodate: concurrent playback of multiple sound sources; moving sound sources; direct and indirect sonification; simpler virtual environments; and complex acoustical models.

Wenzel [1996] outlines that a spatial auditory display ideally meets the following requirements:

1. Adequately reproduce the audible spectrum in frequency resolution and dynamic range

2. Present information accurately in three spatial dimensions

3. Be capable of representing multiple sources, which can be either static or moving

4. Be real-time and interactive; that is, responsive to the ongoing needs of the user

5. Be head-coupled to provide a stable acoustic environment with dynamic cues appropriately correlated with head motion

6. Be flexible in the type of acoustic information that can be displayed; for example, real environmental sounds, acoustic icons, speech, or streams of multidimensional auditory patterns or objects

3.2.1 Binaural auditory display

Binaural auditory display is a subfield of spatial auditory display. Binaural presentation of audio has the advantages and disadvantages provided by headphone playback. While is it portable, it can isolate the user, which depending on the application may or may not be beneficial. The lower cost and relatively simple implementation when compared to loudspeaker systems makes binaural auditory display a popular choice [Wenzel, 1992, Arons, 1993, Schmandt and Mullins, 1995].

Binaural auditory display have a number of implementation decisions which influence the effectiveness of the display. The first is whether the user's own HRTFs are used. Individualised HRTFs are much more difficult to use in a display intended for a large number of people as measuring HRTFs is a time-consuming and costly procedure. However, individualised HRTFs may be available for specialised displays for military or medical applications. Begault et al. [2001] found that non-individualised HRTFs do not degrade localisation, externalisation or reversal rates (rate of front-back confusion) of a speech signal when compared to individualised HRTFs, so non-individualised HRTFs can be effective in the absence of individualised HRTFs. However, other studies with noise sources show improvements in reversal rates when using individualised HRTFs [Wenzel et al., 1993, Moller et al., 1996], so individualised HRTFs are preferred if available, but it is not clear if they provide a direct advantage over non-individualised HRTFs.

Head tracking rotates the sound field to provide a static image when the listener moves their head. Begault et al. [2001] found that head tracking reduces reversals, but it has no effect on localising or externalising a sound source. They also found that individualised HRTFs with head tracking improve azimuth localisation of speech signals. Along with localisation and reversal rates, externalisation is important for creating a realistic virtual sound source. Externalisation is primarily influenced by the presence of reverberation; Begault et al. [2001] found that as little as 80 ms of early reflections is sufficient to make a sound source appear to be outside of the listener's head.

3.3 Virtual Ambisonics

To synthesise a moving sound source with binaural audio, the sound source needs to be convolved with an HRTF at every position along its trajectory. HRTFs are often measured in 5° to 15° increments on the horizontal plane and elevation increments of around 10° to 15°. As see in Table 2.1, the localisation blur for sources directly in front of the listener can be much less than the HRTF measurement increments. If the set of HRTFs is too sparse to simulate the movement smoothly, then interpolation is used to expand the set and synthesise additional HRTFs. This process is repeated for every sound source. If the sound sources are static, but located at a position without a measured HRTF, interpolation is needed. If head tracking is used at all, interpolation is needed even if the sound source or sources do not move relative to the listener. The process to place a sound source in the virtual space is also repeated for every discrete reflection from a room model. The synthesis of a binaural space then quickly becomes computationally expensive, especially as the number of sound sources increases. Virtual loudspeakers are a means to avoid much of the computational cost and can also decouple the cost from the number of sound sources.

Virtual Ambisonics is the rendering of B-format signals over headphones using HRTFs. It was first discussed by McKeag and McGrath [1996], Travis [1996] and has been reiterated in more depth by Noisternig et al. [2003a], Wiggins [2004]. Virtual Ambisonics has large advantages over binaural techniques using only HRTFs, most notably in computational complexity [Noisternig et al., 2003a]. Virtual Ambisonics uses a fixed number of HRTFs independent of the number of sources being rendered and does not depend on interpolation nor a dense measurement set. Virtual Ambisonics requires only as many HRTFs as virtual loudspeakers as seen in Figure 3.1, and as will be developed here, can require even fewer. Computational complexity is further reduced as head rotations that are used in head-tracking systems can be easily implemented by rotating the B-format sound field. These advantages also make binaural audio more feasible on mobile devices

Virtual Ambisonics has the advantages of an idealised listening environment without the problems that are usually presented in setting up a large array of loudspeakers in a room. When Ambisonics is rendered binaurally, the listener's head is kept in the ideal sweet spot; the virtual loudspeakers are easily movable and theoretically can be placed anywhere. For below 700 Hz at first order, 1.9 kHz for second order, and 2.5 kHz for third order, the sound field is accurately reconstructed for the listener. Above these frequencies the area where the sound field is reconstructed is smaller than the average human head [Bertet et al., 2007]. Unlike with physical loudspeakers, there is not a concern that the

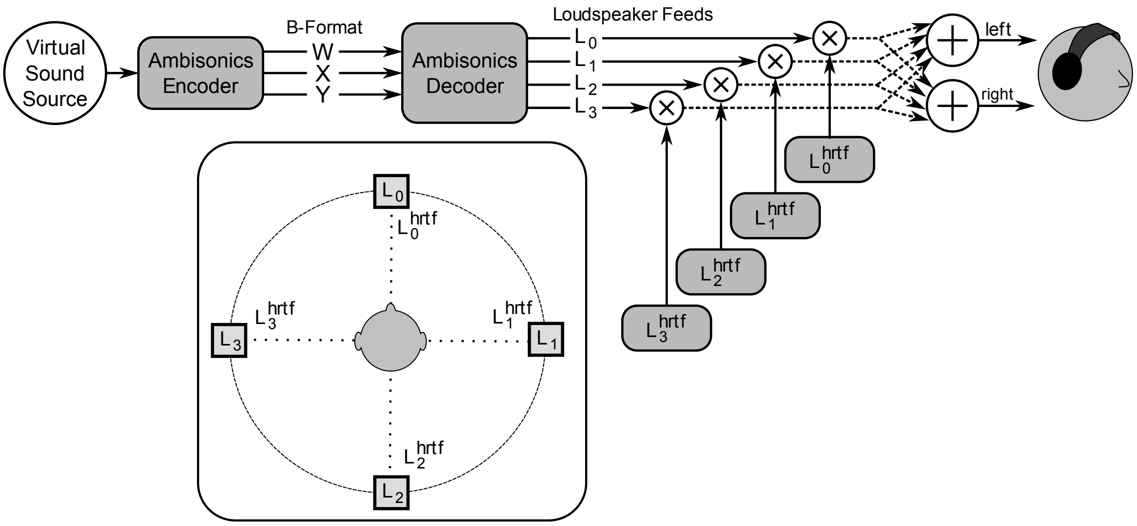


Figure 3.1: General signal flow for binaural rendering of a B-format signal. The virtual loudspeakers could be in any arrangement and are not limited to the positions shown here.

listener may unintentionally move outside the sweet spot. As loudspeaker placement in a regular array can be very difficult to achieve in real rooms, this can be done in the virtual room easily allowing for optimum decoding [Gerzon, 1973]. In practice, the virtual loudspeaker placements are limited by the HRTF set being used. For example, HRTFs are not usually measured directly below the listener, so a virtual loudspeaker could not be placed there.

This thesis examines two special cases of binaural auditory display: auralization of a room model and exploration of a music collection. The broader field of auditory display encompasses a vast array of topics and applications, but with the common goal of conveying information via audio. Auralization and music discovery are both dependent on immense amounts of audio signal processing which is often not necessary in other tasks such as word processing or file management. For this reason, special considerations are needed in regards to rendering the audio information. Virtual Ambisonics is a flexible platform for rendering a binaural signal which will be shown easily adapts to both use cases. However, further optimisation and evaluation is needed to determine whether it performs as well as convolution with HRTFs. Only first order has been optimised in the literature [Wiggins, 2004], though second and third orders are used [Noisternig et al., 2003a]. The different encoding and decoding orders have not been compared with each other or with measured HRTFs.

Only horizontal Ambisonics are examined here. This means there is no vertical information; the elevation of any source will be always equal to zero. This simplifies the encoding and decoding process, but the methods described here could be extended to include height information. This study is also limited to first through third orders.

In the following section the optimisation for first order decoding is derived and then extended to the second and third order. The best practice for virtual Ambisonics then reviewed and the remaining parameters that have not been previously discussed are

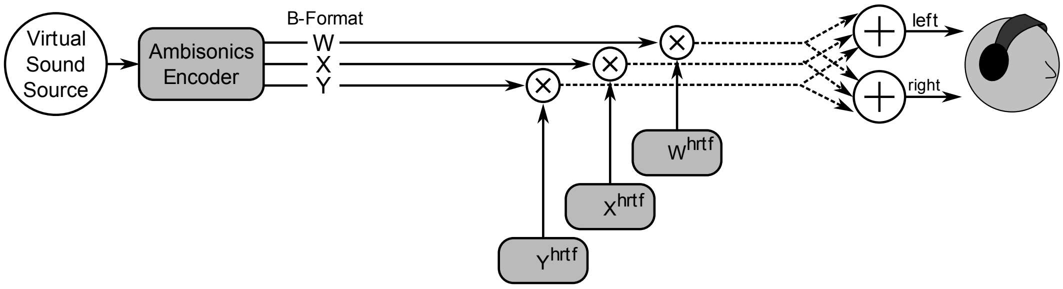


Figure 3.2: General signal flow for binaural rendering of a B-format signal using optimised HRTFs for each Ambisonics channel.

evaluated.

3.3.1 Optimising first order binaural decoding

Though binaural decoding for first order horizontal Ambisonics has been discussed in [McKeag and McGrath, 1996, Wiggins, 2004], it is reviewed here and then extended to the second and third order horizontal decoding. See Section 2.4.3 for a more detailed discussion of Ambisonics decoding.

Virtual Ambisonics requires at least three channels of B-format audio which are mixed down to two channels. A less efficient approach to creating the binaural signal is to compute the virtual speaker feed for each individual loudspeaker and then convolve that loudspeaker feed with the appropriate HRTF, as seen in Figure 3.1. Instead, the decoding can be optimised by finding the HRTF pair for each B-format channel instead of each individual virtual loudspeaker feed. This reduces the number of HRTFs needed and the number of convolutions required to create the binaural signal. So at first order, only three pairs of HRTFs (six filters) are required for any loudspeaker arrangement [Wiggins, 2004]. This translates to six convolutions.

The HRTF for each B-format channel is computed from the chosen virtual loudspeaker layout.

$$\begin{aligned}
W^{hrtf}_{left,right} &= 1/\sqrt{2} \times \sum_{k=1}^{N} \left(L^{hrtf}_{k} \right) \\
X^{hrtf}_{left,right} &= \sum_{k=1}^{N} \left(\cos(\theta_k) \times L^{hrtf}_{k} \right) \\
Y^{hrtf}_{left,right} &= \sum_{k=1}^{N} \left(\sin(\theta_k) \times L^{hrtf}_{k} \right) \qquad (3.1)
\end{aligned}$$

N is the number of virtual loudspeakers each with a corresponding azimuth θ and HRTF, L^{hrtf}. The signals for each ear are are the sum of the B-format convolved with the left and right channels of the HRTFs.

$$\begin{aligned} Left &= (W \otimes W_{left}^{hrtf}) + (X \otimes X_{left}^{hrtf}) + (Y \otimes Y_{left}^{hrtf}) \\ Right &= (W \otimes W_{right}^{hrtf}) + (X \otimes X_{right}^{hrtf}) + (Y \otimes Y_{right}^{hrtf}) \end{aligned} \tag{3.2}$$

As stated by Gerzon [1992], for best results Ambisonics should be decoded to regular loudspeaker arrays. If the virtual loudspeakers are distributed about the listener so that the left and right sides of the listener are symmetric, the decoding can be optimised further. The left and right HRTFs of the omnidirectional channel W are equal, as are the left and right HRTFs of the X channel which captures front/back information. The left and right HRTFs of the Y channel are equal but phase inverted.

$$\begin{aligned} W^{hrtf} &= W_{left}^{hrtf} = W_{right}^{hrtf} \\ X^{hrtf} &= X_{left}^{hrtf} = X_{right}^{hrtf} \\ Y^{hrtf} &= Y_{left}^{hrtf} = -Y_{right}^{hrtf} \end{aligned} \tag{3.3}$$

Now only three individual HRTFs, not the left and right channels of pairs of HRTFs, are needed for a horizontal binaural rendering as seen below [McKeag and McGrath, 1996, Wiggins, 2004].

$$\begin{aligned} Left &= (W \otimes W^{hrtf}) + (X \otimes X^{hrtf}) + (Y \otimes Y^{hrtf}) \\ Right &= (W \otimes W^{hrtf}) + (X \otimes X^{hrtf}) - (Y \otimes Y^{hrtf}) \end{aligned} \tag{3.4}$$

In summary, first order horizontal-only Ambisonic decoding can be accomplished with only three convolutions with three single, not pairs of, HRTFs.

3.3.2 Extending to second and third order decoding

The same optimisations can be applied to second and third order horizontal-only decoding. Second order requires the additional channels U and V, and third order requires P and Q. The HRTF pair for each channel can be computed as had been done for first order by using the appropriate Ambisonics coefficients as seen in Equation 3.5.

$$\begin{aligned}
U^{hrtf} &= \sum_{k=1}^{N}\left(\cos(2\theta_k)\times L_k^{hrtf}\right)\\
V^{hrtf} &= \sum_{k=1}^{N}\left(\sin(2\theta_k)\times L_k^{hrtf}\right)\\
P^{hrtf} &= \sum_{k=1}^{N}\left(\cos(3\theta_k)\times L_k^{hrtf}\right)\\
Q^{hrtf} &= \sum_{k=1}^{N}\left(\sin(3\theta_k)\times L_k^{hrtf}\right) \qquad (3.5)
\end{aligned}$$

The channels U and P share the same symmetries as the X channel; they are symmetrical and in phase. V and Q are similar to Y as they are phase inverted. Equation 3.6 takes these symmetries into account.

$$\begin{aligned}
Left &= (W\otimes W^{hrtf}) + (X\otimes X^{hrtf}) + (Y\otimes Y^{hrtf})\\
&\quad +(U\otimes U^{hrtf}) + (V\otimes V^{hrtf})\\
&\quad +(P\otimes P^{hrtf}) + (Q\otimes Q^{hrtf})\\
Right &= (W\otimes W^{hrtf}) + (X\otimes X^{hrtf}) - (Y\otimes Y^{hrtf})\\
&\quad +(U\otimes U^{hrtf}) - (V\otimes V^{hrtf})\\
&\quad +(P\otimes P^{hrtf}) - (Q\otimes Q^{hrtf}) \qquad (3.6)
\end{aligned}$$

In summary, second order horizontal-only Ambisonics decoding can be accomplished with five convolutions with five single HRTFs and third order can be accomplished with seven convolutions with seven single HRTFs.

3.3.3 Virtual loudspeaker placement

The following criteria have been determined as optimum loudspeaker setups for binaural Ambisonics decoding:

- Regular distribution of loudspeakers [Gerzon, 1992]

- Keep symmetry to the left and right of the listener [McKeag and McGrath, 1996, Wiggins, 2004]

- Use minimum number of loudspeakers required for order [Daniel et al., 1998]

The number of loudspeakers is dependent on the order of the system and it is recommended that the minimum number of loudspeakers should be used to avoid comb-fitering effects from combining multiple correlated signals [Daniel et al., 1998]. A degradation in the binaural signal is found when increasing the number of virtual loudspeakers past the minimum number [Pulkki, 2001b].

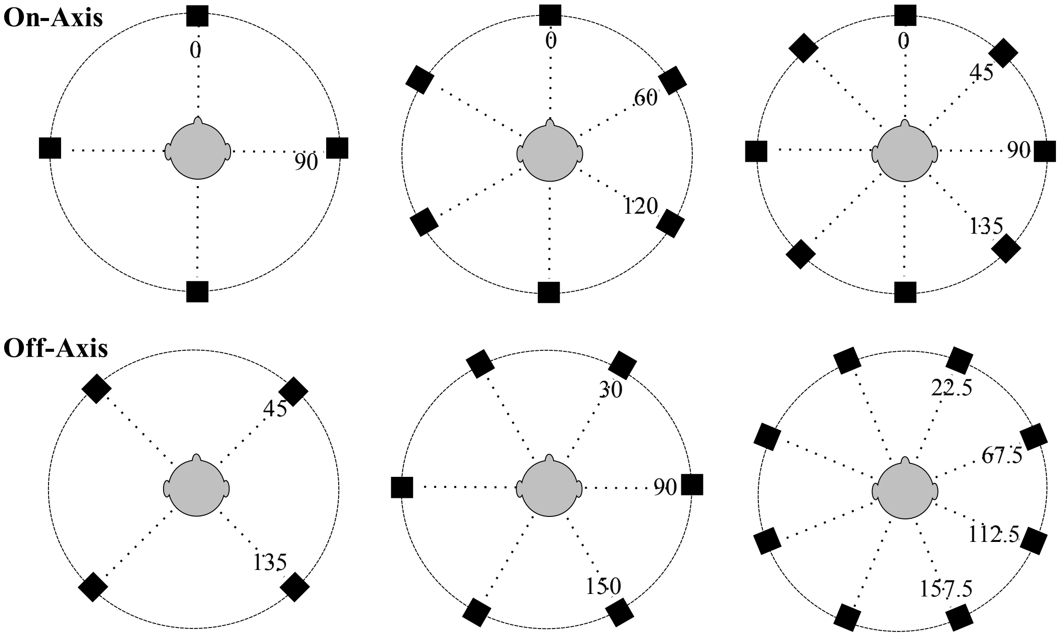


Figure 3.3: Possible symmetric virtual loudspeaker configurations for the 4, 6, and 8 loudspeakers.

$$N \geq 2M + 2 \tag{3.7}$$

where N is the number of loudspeakers required for order M.

These criteria still leave some ambiguity, namely the placement of the virtual loudspeakers. As illustrated in Figure 3.3, this leaves two possible configurations for each order. The second order can have a pair of loudspeakers that are either on-axis with the ears or on-axis with the nose. First and third order have loudspeakers on-axis with both the ears and the nose in the first configuration and off-axis in the second. For third order, there are explicit engineering difficulties in implementing the second configuration as the loudspeakers are placed first at 22.5° and then in 45° intervals. HRTFs are not typically measured at these locations, so without interpolating or synthesising HRTFs, there is no simple way to implement the off-axis configuration.

3.4 Evaluation of Virtual Ambisonics

The evaluation of virtual Ambisonics will determine *if there is a perceptible difference between the binaural signal generated by convolution with HRTFs and virtual Ambisonics.* If there is no perceptible difference or virtual Ambisonics performs better, then virtual Ambisonics can be recommended for binaural auditory display.

To evaluate virtual Ambisonics, HRTFs generated with virtual Ambisonics are compared to the directly measured HRTFs. The evaluation involves signal processing analysis of the HRTFs and listening tests. For the signal analysis, the measured HRTFs are considered to be the 'correct' HRTFs and any deviation from those measurements

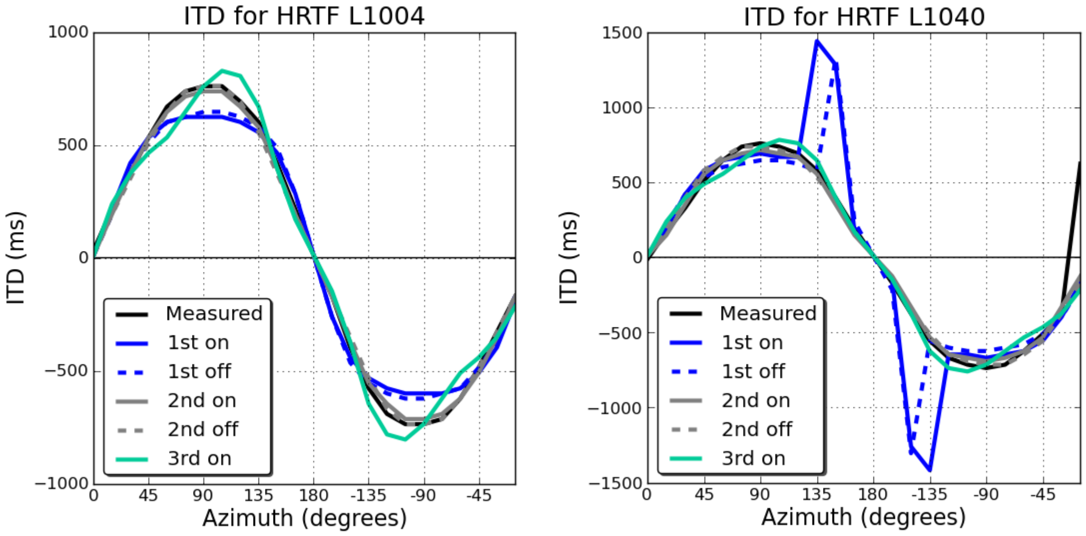


Figure 3.4: The ITD for the original HRTF sets *L1004* and *L1040* and the Ambisonics-derived HRTFs using the same HRTF sets.

will be considered error. Two factors are examined: the loudspeaker arrangements for first and second orders (on-axis and off-axis) and increasing the decoding order from first to third. The signal analysis looks at the ITD, ILD, and spectral error incurred for 24 azimuth positions at 15° increments with an elevation of 0°. Listening tests examine the localisation error of 13 azimuths with an elevation of 0°.

The evaluation uses 11 sets of HRTFs from the LISTEN data-base[1] which will be referred to as *L1002*, *L1004*, *L1007*, *L1014*, *L1016*, *L1020*, *L1023*, *L1032*, *L1040*, *L1046*, and *L1055*. Each LISTEN HRTFs is measured from a different person and not dummy heads. The complete LISTEN database has 51 HRTF sets, the 11 used here were chosen at random.

3.4.1 ITD

As discussed in Section 2.1, ITD is an important cue for localising sound sources with frequency content below 1 kHz, but also affects localisation up to approximately 1.5 kHz. The frequency-dependent ITD values are calculated from white noise convolved with the HRTFs and then filtered with equivalent rectangular bandwidth (ERB) filters with centre frequencies at 600 Hz to mimic the human auditory system. The ITD is then the time delay between the left and right channel determined by cross-correlation. This is repeated for the measured HRTFs and the HRTFs generated by virtual Ambisonics.

The ITD values for the measured and virtual Ambisonics HRTF sets *L1004* and *L1040* can be seen in Figure 3.4. While there is some variation between HRTF sets, in general all of the sets mimic *L1004*. The on and off-axis second order Ambisonics HRTFs follow the original HRTFs the most closely. The first order decoders have a smaller range and the third order decoders have a slightly larger range than the measured HRTFs. The third order virtual Ambisonics HRTFs also are slightly out of phase with the measured

[1] http://recherche.ircam.fr/equipes/salles/listen/

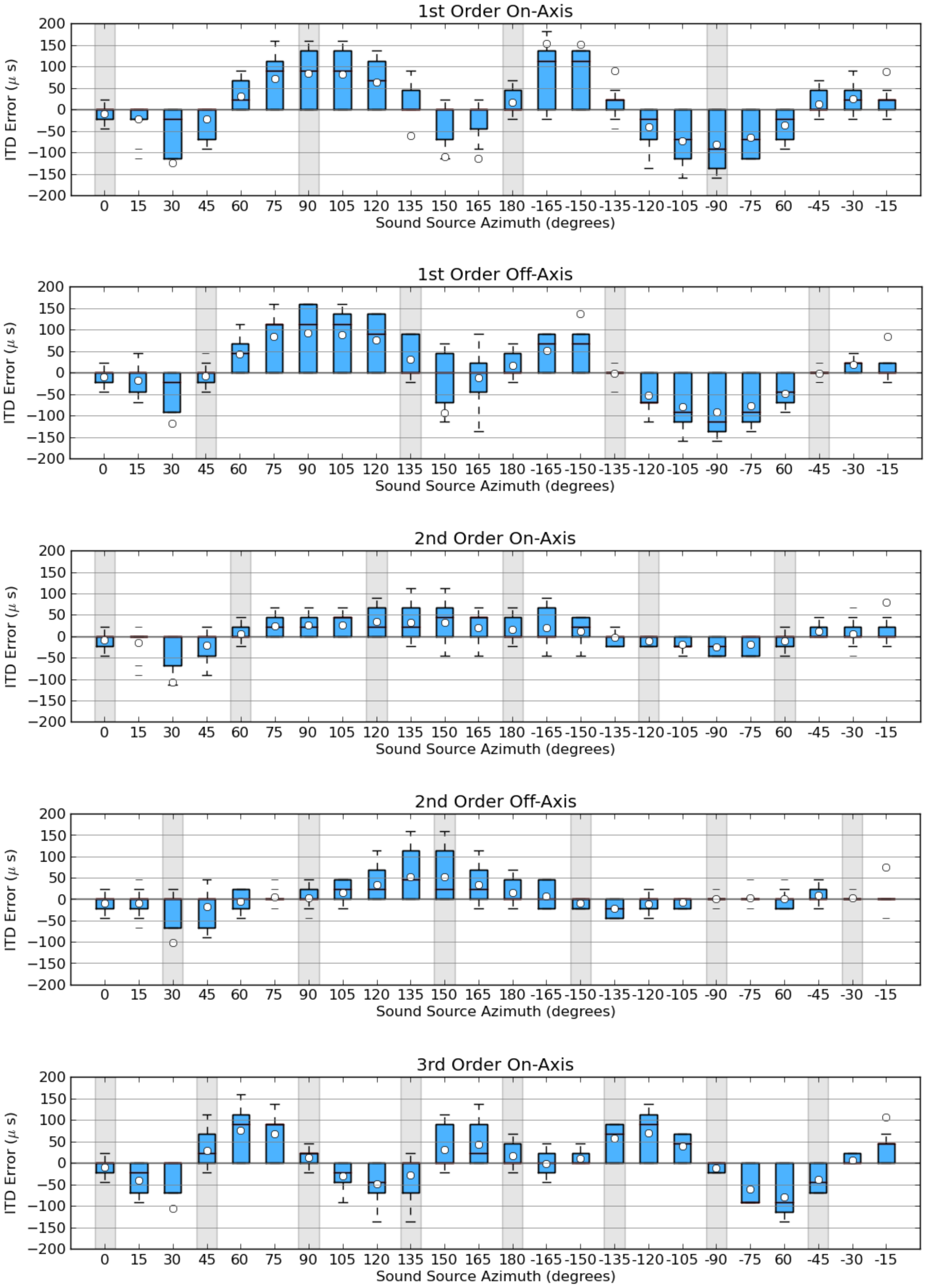


Figure 3.5: Box plots of the difference between the ITD of the measured HRTF sets and the virtual Ambisonics HRTF sets for 24 azimuth positions. The circles indicate the mean.

HRTFs. The maximum and minimum ITD values of the third order occur closer to 180° than the measured HRTFs which have maximum and minimum at −90° and 90°.

There are a few cases of discontinuities in the ITD values that differ greatly from the measured HRTFs as can be seen in the first order virtual Ambisonics HRTFs for *L1040* in Figure 3.4. These discontinuities in the first order virtual Ambisonics HRTFs only occur in three of the eleven HRTF sets rendered with virtual Ambisonics (*L1007*, *L1040*, and *L1046*). These discontinuities all occur at 15° or 30° within the azimuth position 180°. There is an additional discontinuity in the measured HRTF set *L1023* at −90° and −15° (not in the virtual Ambisonics rendered HRTFs), in the measured set *L1040* at −15 (not in the virtual Ambisonics rendered HRTFs), and very irregular patterns in the measured and virtual Ambisonics sets *L1007*. It should be noted that for the listening tests described in the next section, none of the participants selected *L1007* as their preferred HRTF set. These discontinuities can also be observed in the virtual Ambisonics HRTFs evaluated in [Wiggins et al., 2001].

Figure 3.5 shows box plots of the difference in ITD values between the measured HRTFs and the HRTFs synthesised with virtual Ambisonics. The three HRTF sets with large discontinuities in the first order virtual Ambisonics sets are treated as outliers and are not included in the plots. The box plots show the lower quartile, upper quartile, median, and mean of the difference between the ITD of the measured and synthesised HRTFs for each source position across all HRTF sets (excluding the three outliers). The range of error is similar for each virtual Ambisonics decoding method, and the error is extremely similar for the on and off-axis configurations for each decoding order. The greatest difference is when comparing the different orders to each other. The error varies as a function of the source position in a sinusoidal pattern. For each order there is less error at 0° and 180° than to either side. At second and third order decoding the error decreases at −90° and 90°, but first order has its greatest error these positions.

The second order decoding with on or off-axis virtual loudspeaker configurations performs the best as the mean difference in ITD between the measured and synthesised HRTFs is 26.83 μs with off-axis virtual loudspeakers, which is lower than all other orders and virtual loudspeaker configurations.

Even though the measured HRTFs are not symmetrical to the left and right of the listener, the virtual Ambisonics HRTFs are symmetrical. The head is assumed to be symmetrical in the Ambisonics model, though it is not in reality. This is reflected in Figure 3.5 as the absolute error is not symmetric around 180° for any of the virtual Ambisonics decoders, though the errors to the left and right are similar.

3.4.2 ILD

ILD is a complementary cue to ITD as it is an important cue for localising sound sources with frequency content above 1.5 kHz. See Section 2.1 for more details. The frequency-dependent ILD values are calculated from white noise convolved with the HRTFs and then filtered with ERB filters with a centre frequencies of 10 kHz. The ILD is then the difference in decibels between the left and right channel. This is repeated for

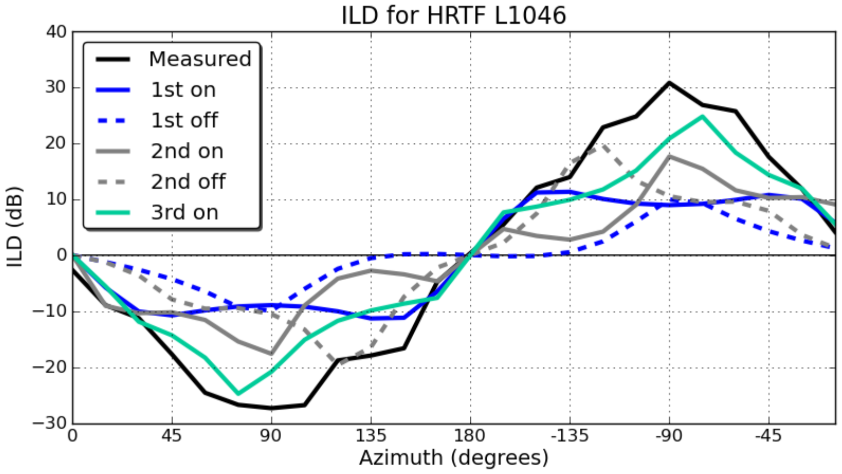


Figure 3.6: The ILD for the original HRTF sets *L1046* and the Ambisonics-derived HRTFs using the same HRTF set.

the measured HRTFs and the HRTFs generated by virtual Ambisonics.

The ILD for *L1046* can be seen in Figure 3.6. The ILD for all of the other HRTF sets are similar to the figure – as the virtual Ambisonics order increases, the ILD cues are better approximated. Wiggins et al. [2001] also found this to be true in their analysis of virtual Ambisonics at the first and second order.

Figure 3.7 shows box plots describing the difference in ILD values between the measured HRTFs and the HRTFs synthesised with virtual Ambisonics. The box plots show the lower quartile, upper quartile, median, and mean of the difference between the ILD of the measured and synthesised HRTFs for each source position across all HRTF sets. As the order increases from first to third, the ILD can improve by as much as 10 dB. The on-axis virtual loudspeaker configuration performs better than off-axis for first order decoding, but worse for second order.

As was true for the ITD, the least error for all decoding orders and loudspeaker configurations occurs at 0° and 180°. For all orders and configurations the greatest error occurs around −90° and 90°. The loudspeaker configuration has little effect on the ILD error. The asymmetries in the ITD error also occur in the ILD error. The error is not perfectly symmetric about 180°, but is approximate.

3.4.3 Spectral error

Figure 3.8 shows the mean spectral differences between the synthesised and measured HRTF at −90° across all of the HRTF sets. Virtual Ambisonics largely increases the magnitude of the frequency response of the contralateral ear (farther ear from the sound source) and has less effect on the ipsilateral ear. The concern by Daniel et al. [1998], Pulkki [2001b] was that the multiple highly-correlated signals due to a large number of virtual loudspeakers would cause comb filtering. Some comb filtering effects can be seen in the upper frequencies as the sound source approaches −90° or 90°, but these effects are not very pronounced. The magnitude ‘smoothes out’ slightly with higher orders, but

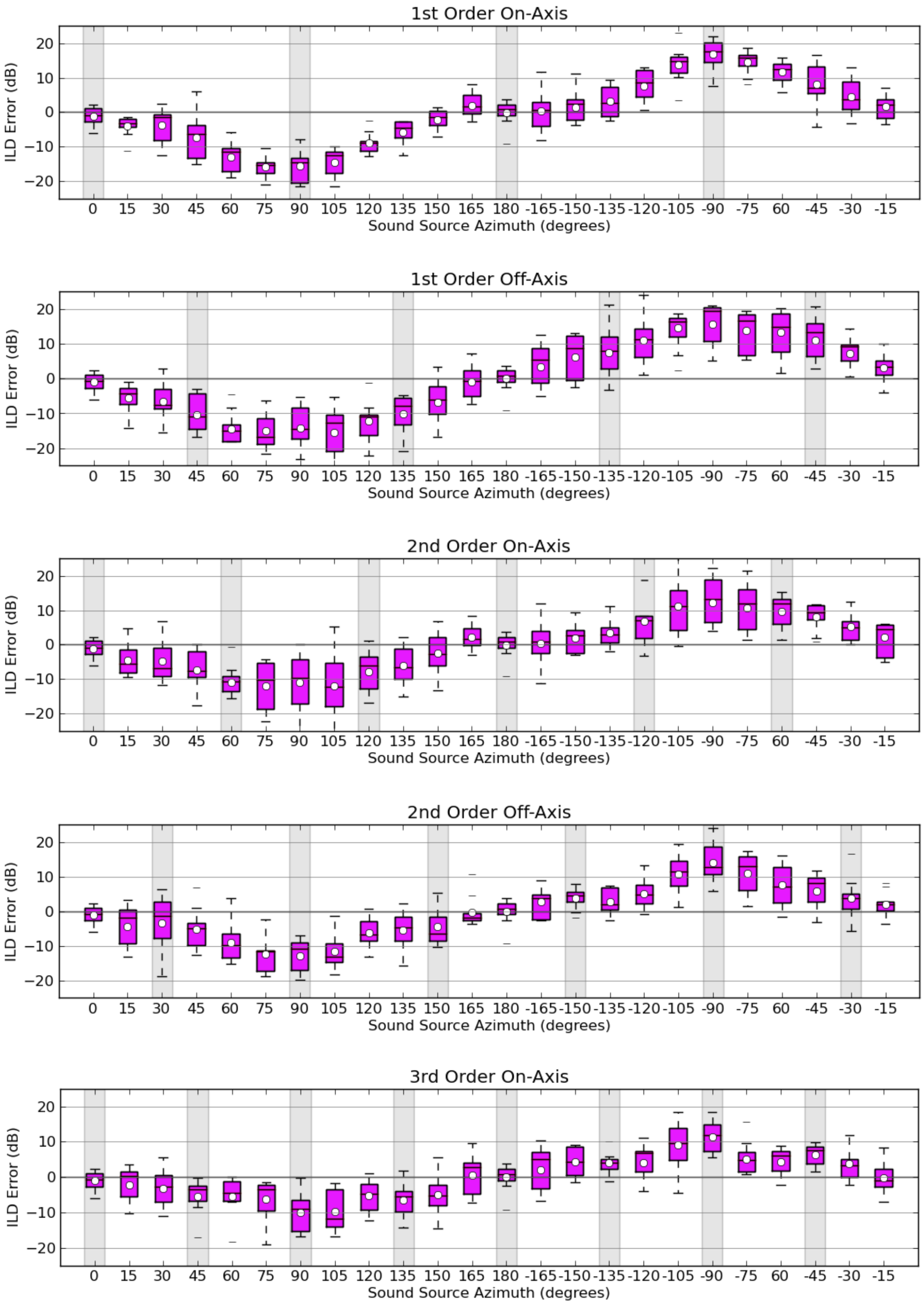


Figure 3.7: Box plots of the difference between the ILD of the measured HRTF sets and the virtual Ambisonics HRTF sets for 24 azimuth positions. The circles indicate the mean.

has no change with regards to the virtual loudspeaker configuration.

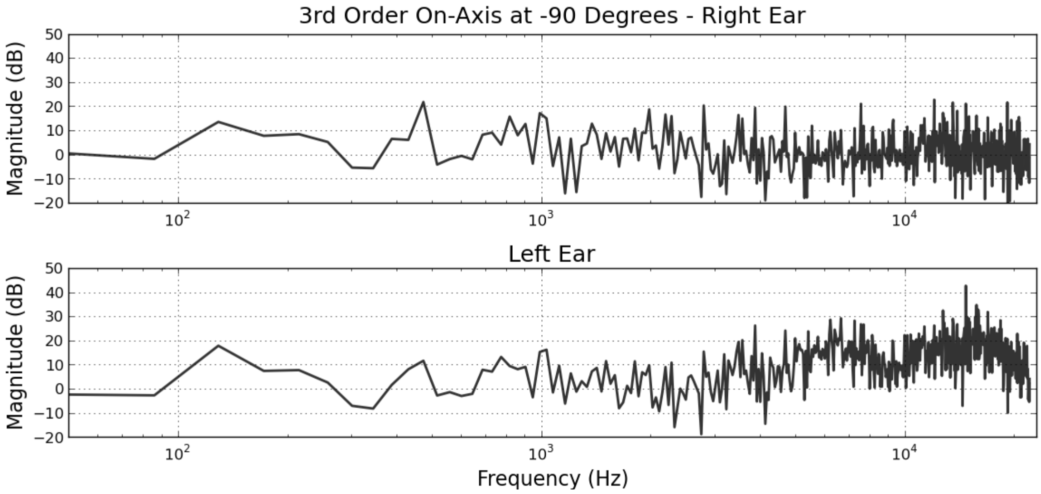


Figure 3.8: Frequency error for the each ear at 90° to the left.

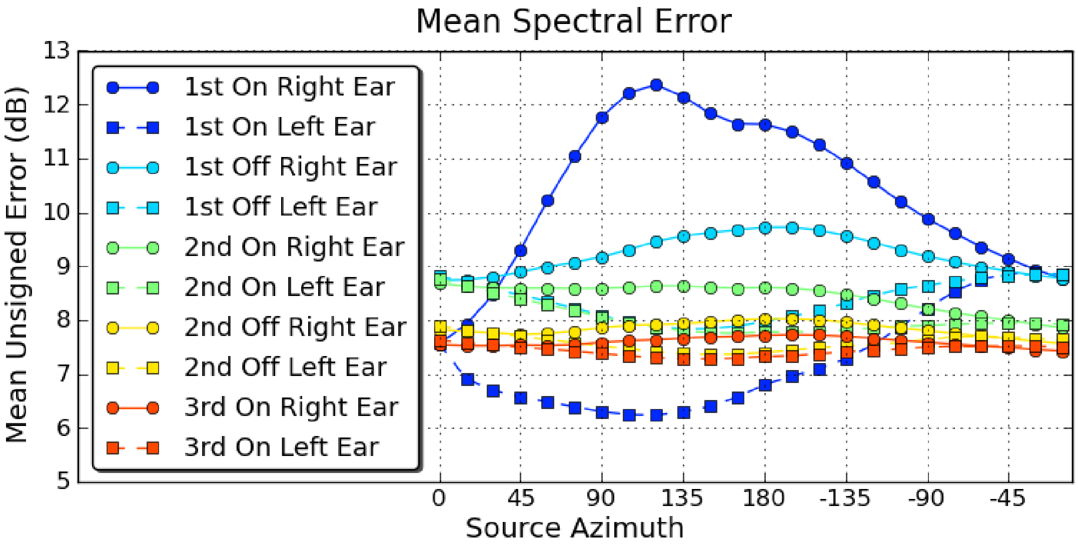


Figure 3.9: Mean unsigned spectral error for all HRTFs.

Figure 3.9 plots the mean unsigned spectral error averaged across all HRTFs for each decoder and source azimuth position. First order on-axis strongly differentiates itself from the other decoders as the right ear has more error and the left ear has significantly less error than the other decoders including first order off-axis. It is also the least symmetric about 180°. This is most likely an artefact of assuming a symmetric head. The HRTFs chosen for the decoders used only the left channel from the measured HRTF and assumed that the right would perform identically to the left. Since the HRTFs were measured from humans with naturally asymmetric heads, the right ear incurs more error. Otherwise, as the decoder order increases, the error decreases. Off-axis virtual loudspeaker configurations also have less error than on-axis for both first and second order.

3.4.4 Listening tests

While virtual Ambisonics creates HRTFs that differ from the originally measured HRTFs, listening tests are required to determine whether these differences result in perceptible errors. The most pertinent error is whether virtual Ambisonics increases the localisation error in comparison with a binaural signal without virtual Ambisonics. If virtual Ambisonics does increase the localisation error, then it should only be used in applications where the increased error can be tolerated. However, if virtual Ambisonics does not introduce additional error, then there it can used in virtual acoustics displays to decrease computational load without degrading the localisation.

Listening tests were conducted with 13 participants, 3 female and 10 male. All participants reported having normal hearing and all but two participants identified as an expert listener. The instructions presented to each participant and the complete set of results can be found in Appendix A.

The test involved two steps: selection of a non-individualised HRTF set from the 11 sets used in the signal analysis and localisation of virtual sound sources. The sound source for both portions of the listening test was a train of three 50 ms white noise bursts separated with 70 ms of silence. The onset and offset of the noise bursts were shaped by the leading and falling edge of a 4 ms Hann window. Noise bursts are commonly used as the sound source for localisation studies [Wenzel et al., 1993, Wersényi, 2009].

Localisation error significantly decreases when participants are allowed to select the best HRTF set from a collection of HRTFs [Seeber and Fastl, 2003], so the first portion of the listening allowed the participant to choose an HRTF set from the 11 available. The noise train was convolved with each HRTF at $-45°$ to $45°$ in $15°$ increments so that the source appeared to be moving from 45° to the left to 45° to right and then back to the left. The participant was asked to select the HRTF that best met a list of criteria such as: the source was a constant distance outside the participant's head; smooth movement of the sound; and the source remained on the horizontal plane. The selected HRTF set was then used in the localisation task. This proved to be a useful exercise as it provided a training phase which allowed participants to become familiar with the sound source, and there was not unanimous consensus on the preferred HRTF set. From the available 11 HRTF sets from the LISTEN database 6 HRTFs were preferred by at least one participant: *L1002*, *L1007*, *L1014*, *L1016*, and *L1020* were not chosen by any of the participants; *L1023* and was chosen only once; *L1004*, *L1032*, *L1040*, and *L1055* were each chosen by 2 participants; and *L1046* was chosen by 3 participants.

In the second task of the listening test the participant was asked identify where a virtual sound source appeared to be located. The HRTF set selected in the first task was used to generate the binaural audio in the second. The perceived position was recorded through a graphical user interface; the participant rotated a dial so that it pointed towards the sound source. The participant could not indicate distance and was forced to choose a location; they could not declare that they did not know.

Each participant was presented with 78 trials. One trial consisted of a single sound source synthesised with one of the six encoding methods (measured HRTFs or virtual

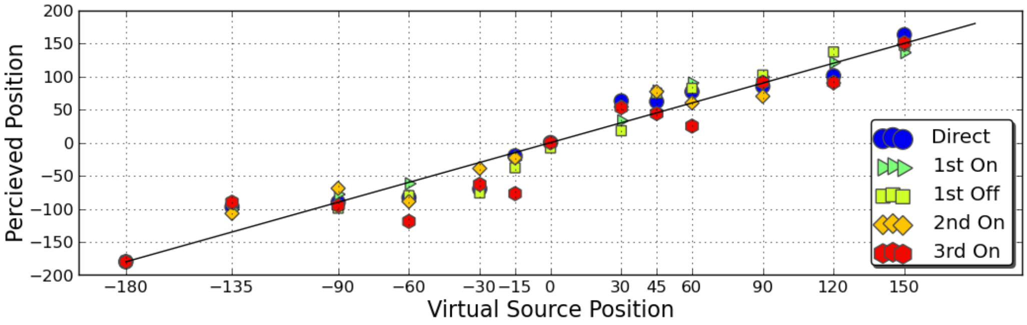


Figure 3.10: The perceived source position for a single participant. The line shows the position of the virtual source. The greater the distance between each marker and the line, the greater the error.

Ambisonics) at one of 13 positions on the horizontal plane. To evaluate the asymmetrical error introduced by assuming a symmetrical head, the virtual sound source was placed at 30°, 60°, and 90° to the left and right of the listener. The other seven positions were: 0° and 180°; 120° and 150° to the right; and 15° and 135° to the left. The tests were double blind, neither the person taking nor administering the test knew how or where the source was synthesised, and the order of the trials was randomised for each participant.

3.4.4.1 Azimuth error

The azimuth error is the difference between the virtual sound source location and the perceived location. Figure 3.10 shows the perceived azimuth for 78 trials performed by the same listener. Any front-back confusions were corrected by moving the indicated location for a sound source to the correct hemisphere. The different markers indicate where the source was perceived to be located according to each encoding method. As this data alone does not conclusively indicate which encoding method is most accurate, Figure 3.11 shows the average unsigned error for each participant and encoding method. The unsigned error is the absolute value of the difference between the markers in Figure 3.10 and the intended location (the line in the figure). The azimuth error found in these tests corresponds with the error found in other binaural localisation tests such as Begault et al. [2001].

Figure 3.11 shows which encoding method resulted in the least mean unsigned error over all of the source positions for each participant. There is no clear consensus on which encoding method is performed best. When considering all encoding methods including direct convolution with the HRTF, 23% of the participants had the least error with the direct method, 23% had the least with first order with on-axis virtual loudspeakers 13% with off-axis; 13% had the least with second order on-axis and 23% with off-axis; and none of the participants had the least error with third order. An ANOVA analysis of the mean unsigned error for each participant found that it cannot be concluded that there is a significant difference between any of the methods ($f = 0.3085, p = 0.9062$). This means that this listening test found that *all of the virtual Ambisonics methods perform no better nor worse at localising a broadband sound source than direct convolution with*

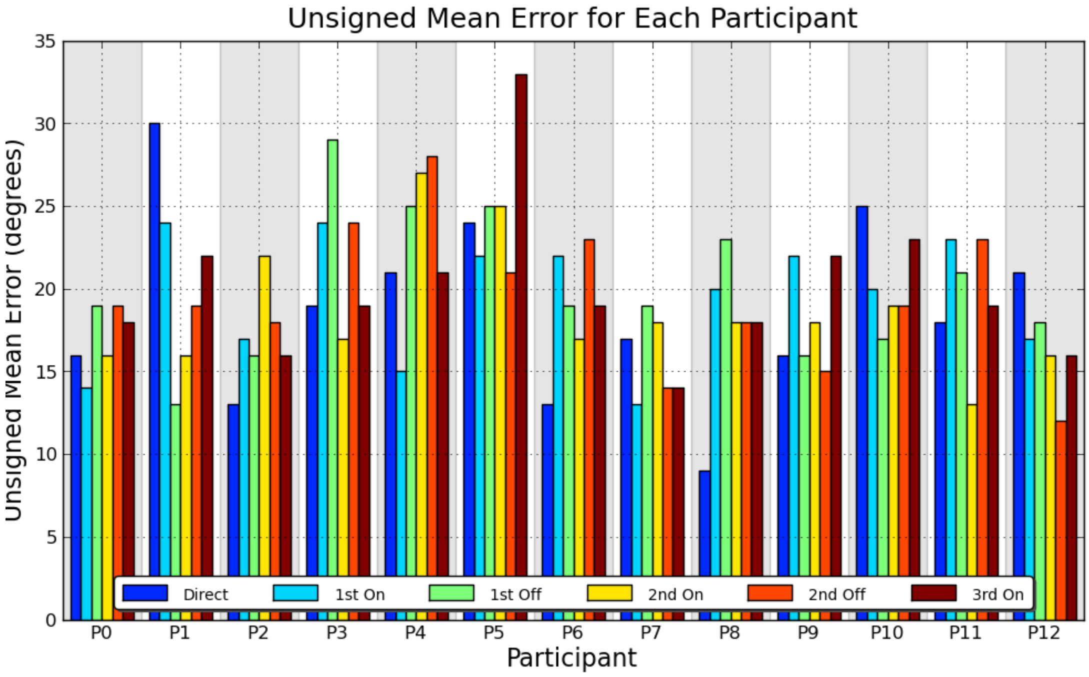


Figure 3.11: The mean unsigned error for each encoding method for each participant. The more perceptually accurate the encoding method the lower the mean error.

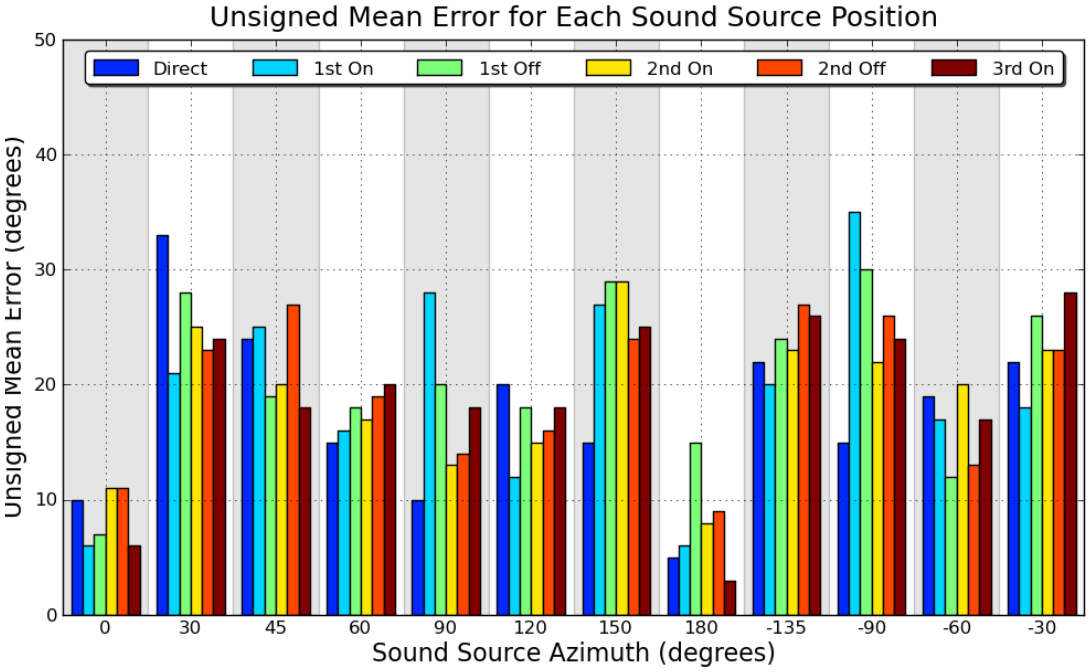


Figure 3.12: The mean unsigned error across all participants for three sound source positions.

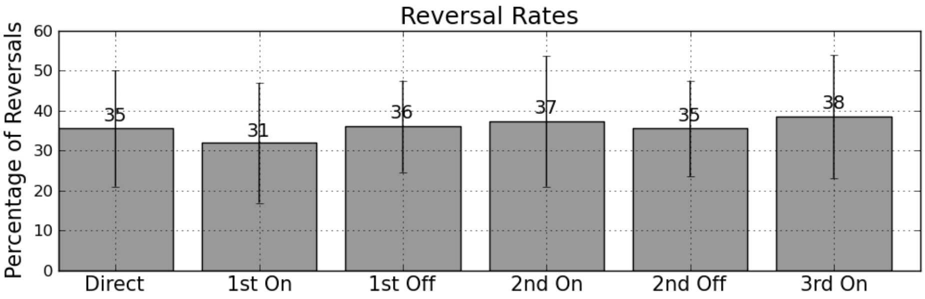


Figure 3.13: The mean reversal rates for each binaural encoding method. The error bars show the standard deviation.

an HRTF. Additionally, as seen in Table 2.1, the localisation blur for broadband noise is 3.2°; many of the differences in error between encoding methods for individuals is within the localisation blur, further showing that is no perceptible difference between methods.

As the HRTF sets used in the listening test are measured from real people, not dummy heads, the HRTFs are not symmetrical to the left and right. However, the virtual Ambisonics encoding methods assume that the head is symmetrical. This assumption may increase the amount of error incurred for sources positioned to one side versus the other. Figure 3.12 shows that there is not a consistent difference in the error when sources are the same angle to the left and right. The error at −30° and 30° and −60° and 60° are within 5° of each other, but the difference in the error at −90° and 90° approaches 15°. This implies that the symmetry assumption of virtual Ambisonics does not greatly influence perceive localisation error, but cannot be disregarded. Further study is needed to determine how prevalent this difference in error for more azimuth positions and types of sources.

3.4.4.2 Front-back reversals

Front-back reversals are when a sound is misinterpreted as being located behind or in front when the opposite is true. They are amongst the most common errors with binaural audio and can only be diminished with head tracking [Begault et al., 2001]. The listening tests did not use head tracking, so the mean reversal rates for each encoding method shown in Figure 3.13 can be considered the maximum rates. A system with head tracking should exhibit lower reversal rates. The rates observed are very similar to those found in [Begault et al., 2001].

An ANOVA analysis of the mean reversal rates for each encoding method does not show any significant difference between the methods ($f = 0.1799, p = 0.9694$). This means that it cannot be inferred from this data that virtual Ambisonics has any effect on front-back or back-front reversals significantly different from direct convolution with an HRTF.

3.4.5 Discussion

This evaluation looks at how the virtual loudspeaker placement and the decoding order affects the binaural rendering of horizontal-only Ambisonics and whether the effects are perceptible. A summary of the evaluation can be seen in Table 3.4.5. In short, the signal

Order	**Loudspeaker Configuration**	**ITD** (μs)	**ILD** (dB)	**Spec. ME** (dB)	**Az. Error** (°)	**Rev. Rate** (%)
1st	On-axis	72.56	8.08	10.34	19.35	31
	Off-axis	60.66	9.66	9.22	20.48	36
2nd	On-axis	29.48	7.28	8.43	19.29	37
	Off-axis	26.83	6.95	7.84	19.89	35
3rd	On-axis	44.78	5.84	7.59	20.25	38

Table 3.1: Means of the error measurements for each Ambisonics order and virtual loudspeaker configuration. The mean spectral error for only the right ear is listed as it is greater than the left ear.

analysis shows that there are measurable differences between the measured HRTFs and those synthesised with virtual Ambisonics, but listening tests show that these errors do not affect localisation nor reversal rates.

The differences between the measured and synthesised HRTFs are dependent on the sound source azimuth and order of the decoder. The arrangement of the virtual loudspeakers has little effect on ITD, ILD, or the frequency response. Off-axis configurations tend to perform a slightly better than on-axis. In general, the higher the decoding order, the more accurately the HRTFs are reproduced. The exception is with second order decoding which recreates ITD cues better than first or third order decoders.

Ambisonics differs from panning spatialisation algorithms such as VBAP by constantly feeding signal to all loudspeakers regardless of the location of the virtual source. When using stereo panning, sources located closest to one ear (panned completely left or right or located at 90°), no signal is feed to the opposite ear. However, in virtual Ambisonics and convolution with an HRTF the contralateral ear will still receive some amount of signal. With virtual Ambisonics, the amount of signal is dependent on the decoding order; theoretically, the higher the order the more precise the placement of the source in the virtual sound field. This is seen in all aspects of the signal analysis as the error increases as the sound source moves away from the median plane and towards either ear. This is especially evident in the ILD and spectral errors.

It is also important to note that not all HRTFs react the same way. Eleven HRTF sets were used the evaluation because not all HRTFs exhibit the same errors when used in virtual Ambisonics. Figures 3.5 and 3.7 show the variance of the error for all HRTFs. There are common patterns in the error that emerge in the ITD, ILD, and spectrum, but the averaged results from the signal analysis found here may not be directly applicable to all HRTFs.

Listening tests were performed with 13 participants. Individuals perform better with a particular decoding order and loudspeaker configuration, but when to generalising to a larger population, no single method gives a significant advantage over any other.

Bertet et al. [2007] found when performing localisation listening tests to evaluate Ambisonics microphones with that third order did no better than second order, though both were significantly better than first order. The listening tests performed in this study showed no difference between first through third order.

These listening tests studied the localisation error and reversal rate of broadband noise for non-individualised HRTFs without head tracking. While these tests show that reversal rates and localisation errors are not siginificantly different between the approaches, other factors need to be considered. Head tracking and reverberation may reveal differences and should be tested in future studies. These tests did not study externalisation or perceived source width. Assuming a symmetrical head, as virtual Ambisonics does, may cause problems with externalisation [Brookes and Treble, 2005]. Further studies would be needed to investigate this.

3.5 Summary

Auditory displays convey information with audio and can be used to reinforce or complement other modes of interaction such as visual display, or can be the sole interaction mode. The term covers a variety of applications including systems to monitor complex real-time information such as air traffic control or medical monitoring equipment. The virtual environment created by an auditory display can be multi-dimensional and use two or three-dimensional acoustic space to display information. The physical accuracy of this space varies according to the application. A model of a newly designed concert hall needs accurate reproduction of a specific space, while an application to navigate calendar information on a mobile phone may create a space that would never exist in reality.

The signal processing requirements of spatial auditory display are dictated by common design approaches. These include: more than one sound source; concurrent playback of multiple sources; moving sources; and virtual acoustics varying from simple to complex room models. Binaural auditory display has further considerations such as whether to use head tracking, reverberation, and individualised versus non-individualised HRTFs.

A flexible spatialisation algorithm is needed to accommodate a generic binaural auditory display. Convolution with HRTFs does not scale well as the number of sound sources increases and room models add complexity. Movement of sources or head tracking greatly increases the computational load as HRTF interpolation is needed. Just two sources requires four convolutions with HRTFs or only three convolutions with virtual Ambisonics. Movement further requires convolution with an additional two sets of HRTFs so that cross-fading between the current and new positions of the virtual sources can take place. Interpolation may also be necessary. Movement with virtual Ambisonics occurs in the B-format domain, so no additional convolutions are required. Additionally, rotation of the entire sound field requires only a matrix multiplication.

Virtual Ambisonics refers to the binaural decoding of a B-format signal by convolving virtual loudspeaker feeds with HRTFs to create a binaural signal. It does not limit the number of sources nor complexity of the room model as the number of convolutions is independent of the number of sources. The technique has been used in previous work as an efficient means to render a binaural signal, but it has not previously been thoroughly evaluated. In particular, the effect of the placement of the virtual loudspeakers in horizontal-only reproduction and the effect of increasing the order of an optimised decoder had not been examined.

To evaluate virtual Ambisonics 11 HRTF sets were used to synthesise new HRTFs using five different virtual Ambisonics decoder parameters: three decoding orders and two different loudspeaker configurations for first and second orders denoted as on and off-axis. On-axis loudspeaker configurations for the first order consist of virtual loudspeakers located directly to the right, left, front, and back of the listener. For the second order the on-axis loudspeaker configuration places virtual loudspeakers directly in front and behind the listener while the off-axis places virtual loudspeakers directly to the right and left of the listener.

The evaluation consisted of comparing the synthesised HRTFs to the measured HRTFs with signal analysis and listening tests. The signal analysis found that generally the synthesised HRTFs better mimicked the measured HRTFs as the decoding order increased and when using an off-axis virtual loudspeaker configuration (when available). However, the listening tests found that any differences in the signal were not perceptible to listeners. Listeners localised a broadband sound source with the same accuracy no matter the rendering technique. However, the listening tests did not examine how virtual Ambisonics may affect other aspects of the spatialisation such as externalisation or timbral changes. Therefore virtual Ambisonics may perceptibly alter the signal in another way, but it does not influence the perceived location on the horizontal plane or reversal rates.

Section 3.2 outlined six general requirements listed by Wenzel [1996] that an auditory display system needs to meet. Virtual Ambisonics has been shown here to meet five of these six requirements. The missing sixth requirement involves head tracking which was not tested here, but virtual Ambisonics could easily accommodate head tracking more efficiently than binaural audio with HRTF interpolation. The other five requirements that virtual Ambisonics has been or will be shown to meet are: adequately reproducing the audible spectrum in frequency and dynamic range; present information accurately in two spatial dimensions (though the full requirement is three dimensions); be capable of representing multiple sources, moving or static; and be flexible in the type of acoustic information displayed. Virtual Ambisonics has adequately reproduced the audible spectrum in two dimensions because participants in the listening test could locate a broadband noise source on the azimuth plane. The subsequent chapters will demonstrate that virtual Ambisonics can represent multiple moving and static sources, and that it can represent simplistic or complex room models.

Chapter 4

A Novel Reverberation Model for Real-Time Auralization

Auralization models of acoustic spaces aim to exactly replicate the acoustic information of the space. This information is captured in an impulse response (IR). An IR contains the acoustic information transmitted from a sound source to a listener or receiver and can also be empirically measured in an existing space using a number of methods (see Section 2.3). A single IR only contains information pertinent to the positions of the source and receiver; another IR is needed to represent any other source and receiver configuration. When convolved with a dry sound source, the measured space is auralized with the sound source appearing to have been recorded in that space. This process usually has static sound source and receiver positions as the sound source is convolved with a single IR.

Room models calculate the signal that is generated by a sound source and is then received at a particular location. How that signal is calculated varies in complexity and, in general, the more complex a model is, the more accurate. Even complex models can be interactive models given enough memory and adequate processor speed, so the concept of a real-time interactive model changes with the advancement of technology. The primary research question then becomes: *what simplifications can be taken to decrease the memory and computational complexity of an acoustics model?*

This chapter proposes an interactive model which uses a large number of IRs that have been rendered offline. The complexity of the model used to create the IRs then has no impact on the interactivity of the model. Real-time interaction is simulated by using the early portions of the IRs which contain information specific to the receiver position with an averaged model of the late reverberation. By using B-format IRs, the room model contains the signal processing advantages of virtual Ambisonics.

Spatial auditory display of room acoustics models are introduced within the context of archaeological acoustics. Section 4.1 overviews the field of archaeological acoustics in regards to its relation to acoustic modelling. The perceptual and processing limitations placed on interactive models are discussed and a solution based on hybrid reverberation is proposed. Sections 4.2 and 4.3 develop and evaluate the hybrid reverberation approach.

	Ray	*Wave (2D)*	*Wave (3D)*	*Empirical*
Single Receiver	00:13	00:37	14:18	1:00
359 Receivers	70:00+	00:37	14:18	48:00

Table 4.1: Comparison of times to calculate one or more IRs based on approach (minutes:seconds). Ray-based and wave-based times based on the acoustics modelling software ODEON and RenderAIR [Murphy et al., 2008] as listed in [Southern et al., 2009]. The empirical times are based on the time to measure IRs with a microphone and loudspeaker for the collection of IRs described in Appendix B.

4.1 Motivation

Interactive models of rooms, such as those commonly used in video game audio, need to output audio in real-time according to the dictated game environment and actions by the user. This includes moving sound sources and/or a moving listener. The constraint to keep the model interactive without any perceived latency limits the accuracy of the model. Table 4.1 highlights some examples of the processing time required to render an IR or a collection of IRs. The time needed for both the empirical and ray-based approaches increases with the number of IRs, whilst the time needed for the wave-based rendering using a finite element grid remains constant regardless of the number of grid points an IR requires. The empirical and wave-based approaches create more accurate IRs [Lokki and Savioja, 2005], but for a single receiver require significantly more processing time than the simpler ray-tracing. Rough estimates of an IR are acceptable in many instances, but special cases may require a more precise IR. One such case is archaeological acoustics.

Archaeological acoustics is a growing research field that explores historic spaces to gain understanding about the acoustics and the subsequent implications about the people that created and used those spaces [Blesser and Salter, 2007, C. 3]. While artefacts from archaeological digs may be extracted, studied, and displayed in museums for the public, the same cannot be done with acoustical environments. Historic spaces may not be publicly accessible due to geographic location, conservation concerns, or the site may be damaged and no longer a complete structure. Recordings may be possible, but these do not allow the public to explore and fully experience the space. The historic space may be modelled in great detail, but current hardware requires simpler models in order to be interactive. Mobile applications also require lower memory and processor loads.

4.1.1 Archaeological acoustics displays

The field of archaeological acoustics has grown in recent years as researchers have begun to discover similar aural phenomena across multiple archaeological sites. Researchers have been working to answer: did previous civilizations have an understanding of acoustics; were these acoustic phenomena intentional design features; did these phenomena have a direct impact and role in the society such as in religious ceremonies? Engineers approach these questions by meticulously measuring the spaces so that the spaces can be auralized and further studied. Recently audio engineers have acknowledged a need to preserve and record spaces of significance such as famous concert halls and sites of historical acoustical significance [Farina and Ayalon, 2003, Murphy, 2006, Abel et al., 2008].

The Workshop on Virtual Audio and Past Environments[1] at the University of York in 2008 addressed many of these issues. Pertinent work was presented by Gui Campos from the University of Aveiro in Portugal. The Painted Dolmen (Anta Pintada) is a neolithic site in Portugal with fragile paintings which have already been significantly damaged in previous archaeological excursions, so it is not open to the public. The Portuguese government wanted to create educational tools so that the public could still experience the heritage site without causing further damage. This is an increasingly popular enterprise for governments, as both the British and Italian governments have funded similar projects [Murphy, 2006, Ardito et al., 2008].

Researchers from the University of Aveiro used a laser scanner to precisely measure the space and then model it for virtual reality simulation. Though the data existed to create a complex, detailed model of the space, it could not be auralized in real-time, so a simplified model was instead implemented. A similar application was developed for an archaeological park in Italy using GPS and custom software for mobile phones [Ardito et al., 2008]. The researchers found that including sounds to recreate the soundscape was well-received by the students that tested the system. However, even though they had three-dimensional models of the ruins, they did not use any auralization, real-time nor previously rendered.

Applications using virtual acoustic environments are being developed for differing end-users with priorities ranging from high-precision acoustic recreation with a lesser focus on interactivity to a large focus on interactivity at the expense of accurate acoustic models. In between is the emerging field of ‘edutainment’ which hopes to use the interactivity of virtual environments to attract and teach students about specific acoustic environments. However, the signal processing is falling short. While great advances are being made in auralizing three-dimensional models of spaces, to the best of this author’s knowledge, complementary technology has not been sufficiently developed to aid in the real-time interaction with this data.

A visual parallel is computer animation. Feature-length films are created in non-real-time by computers by rendering the individual frames offline. In contrast, video games require the game console to produce images as the player moves in the game. The visuals in video games do not look as realistic as animated films, but they are quickly approaching that quality as the hardware improves. The same is true of virtual acoustics, high-quality audio can be rendered offline, but hardware improvements are needed in order for real-time, interactive audio of the same quality to be generated. Even with the availability of faster hardware, low memory and processor loads will still allow lower production costs.

4.1.2 Requirements for real-time interaction

For a system to be considered interactive and real-time, it needs to be responsive enough that a user does not notice any latency. The system needs to update according to user input and output the new information quick enough that it appears to be instant. This

[1] `http://www-users.york.ac.uk/~dtm3/acousticsheritage.html`

Action	Maximum Latency	Filter content to be updated
Source rotation	35 ms	Source directivity
Head rotation	35 ms	Binaural processing, new HRTF
Translational listener or source movement > 0.25 m	700 ms	Binaural processing and any early reflections
Translational listener or source movement > 1 m	3 s	Binaural processing and possibly entire room response depending on model.

Table 4.2: Requirements for a system to be perceived as real-time. Adapted from [Vorländer, 2008, p. 278]

means that there must be a balance between how long the system can take to compute the new information and how long before the user notices the computation time. For virtual acoustics systems, an update rate of 60 Hz and a total delay of 50 ms is acceptable. If there is a corresponding visual display with the audio output, the audio can be delayed up to 30 ms behind the visuals [Vorländer, 2008, p. 268]. The update time can vary according to the source and listener movement as is shown in Table 4.2.

4.1.3 Hybrid reverberation

An increasingly common approach to interacting with a complex, non-real-time acoustics model is to render a large number of IRs offline and then cross-fade between them in real-time to simulate movement [Dalenbäck and Strömberg, 2006, Southern et al., 2009]. This requires the storage of a large number of IRs, but only the first 80 ms or so are unique to each source/receiver position. After the room is mixed, the late reverberation does not significantly change with receiver position. See Section 2.3 for a review of reverberation.

Since the seminal publications by Schroeder [1962] and Moorer [1979], digital artificial reverberators have contained a component intended to create discrete echoes in order to simulate early reflections and a component to create a set of reflections as dense as possible. Later developments with feedback delay networks acknowledge that prior to high frequency attenuation, the reverberator should produce white noise [Jot et al., 1997]. Moorer [1979] first discussed using frequency-shaped Gaussian noise to simulate the energy in late reverberation. The transition from early reflections to late reverberation can then be modeled as a deterministic system that transitions into a stochastic one [Blesser, 2001]. The statistics regarding early reflections are of greater concern here because the statistics of late reverberation has been covered in depth, particularly in [Jot et al., 1997]. Abel and Huang [2006] have also recently explored similar statistics as examined here, looking at measures of reverberation quality, particularly for judging artificial reverberation. Section 2.2.4 provides an overview of the previous work.

Hybrid reverberation models a complete room IR as a deterministic function that transitions into a stochastic one. This is not a novel concept and has been used in reverberation techniques since work by Moorer [1979]. Hybrid reverberators differ

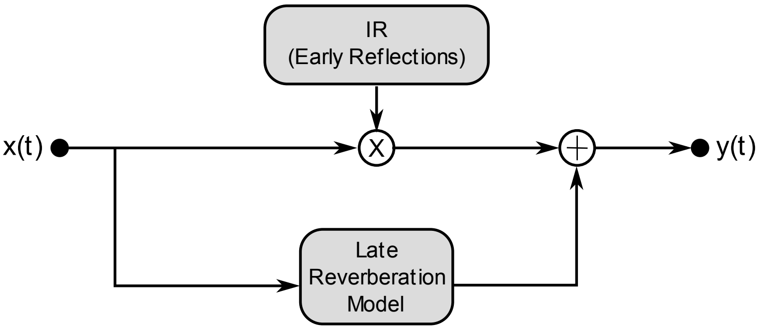


Figure 4.1: Basic construction of a hybrid reverberator.

from traditional statistical reverberators which use a models such as image-source or ray-tracing to calculate the deterministic early reflections. Instead, the early reflections are derived directly from a measured IR and the late reverberation is modeled with a filterbank reverberator. Hybrid reverberation greatly decreases the computational load of convolution reveberation, but theoretically achieves the same results. This technique of combining convolution reverberation with filterbank or recursive reverberation was first explored in [Browne, 2001, Radford, 2003, Stewart and Murphy, 2007], but is extended here.

Hybrid reverberation algorithms rely on two questions: *what portion of the IR should be used for convolution reverberation and what technique will model the late reverberation?* Section 4.2 will look at using kurtosis to determine what portion of the IR to use. Section 4.3 will examine using convolution reverberation also for the late portion of the reverberation. Semantically speaking, the algorithm is no longer hybrid as a convolution is used for the entire IR, but the method is still based on previous work on hybrid reverberation.

Only the case of a moving listener with a static source is examined here as an initial approach and simplification of the problem. This technique could be extended to a moving source with static listener and then moving source and listener in further work.

4.2 Manipulation of Early Reflections

As was discussed in Section 2.2, an IR of a space consists of the direct sound, early reflections, and late reverberation. The early reflections are a set of discrete reflections whose density increases until individual reflections can no longer be discriminated and/or perceived. While the direct sound is a single event that can be easily identified, the early reflections and late reverberation of an IR are more difficult to distinguish. The transition or mixing time of an IR is the earliest point in time when the density of the reflections has reached a statistical threshold in which individual reflections have formed a noise-like signal.

4.2.1 Identifying the direct sound and early reflections

In the literature the transition from early reflections from late reverberation is either defined by a point in time regardless of the room properties, such as 80 ms, or is calculated based on physical properties of a room, most commonly volume. Standardised measurements of room acoustics divide the IR of a room into an early and late portion in order to calculate early lateral energy, clarity and definition. In most literature, it is accepted that the early reflections are contained within the first 80 ms [Begault, 2000]. ISO standards also define the early portion of an IR as the first 80 ms after the impulse arrives at the listening position (when evaluating an IR for musical signals, 50 ms for speech signals) [ISO 3382, 1997]. Section 2.2.4 discusses additional metrics to determine the transition point from early to late reverberation.

Hybrid reverberation transitions from the IR to the statistical reverberator ideally after the early reflections in the IR. Since convolution is more computationally expensive than a recursive filter, only the portion of the IR containing the direct sound and early reflections should be convolved and the remaining portion of the IR should be modelled with the statistical reverberator. A static time such as 80 ms is insufficient as it does not consider the signal content of the IR. The early reflections may be prematurely truncated or excessive late reverberation signal may be convolved.

Unlike the energy from early reflections, the late reverberation of an IR tends towards a normal distribution. A progression towards a more normal distribution occurs as time increases and the acoustic energy within the space becomes more mixed [Blesser, 2001]. A measurement of distribution can then be used to determine whether a point in time is more or less deterministic. That is, whether it is within the early reflections or late reverberation.

The direct sound of an IR needs to travel the physical distance separating the source and receiver. IRs typically begin when the impulse is emitted from the sound source, so there is a delay proportional to the distance between the source and receiver. This delay provides important psychoacoustical cues especially when a multichannel IR is recorded as the relative delays between the channels need to be preserved; the relative delays between multiple measurements also need to be retained. However, convolution is not necessary to create a delay in a signal when a simple delay line will suffice. For hybrid reverberation, the delay is removed from the IR before the mixing time is determined. The delay is reintroduced during synthesis.

If there is no obstruction between the source and receiver, it can be assumed that the first and loudest sound to arrive at the receiver is the direct sound.

$$\tau = \text{argmax}(h(t)) - 0.3 \tag{4.1}$$

The length of the delay, τ, is 0.3 ms before the absolute maximum of the signal. The value of 0.3 ms is arbitrary and is used to ensure that the delay does not include any portion of the direct sound.

The early and late portions of an IR are defined by the mixing time. The mixing time is the earliest point in time that the IR can be considered diffuse and adequately represented by a statistical model. Selecting the mixing time to be a single point in time without considering the space may result in an inaccuracy, but access to the dimensions of a space may be impractical or impossible. A blind method that can determine the transition point between early reflections and late reverberation without knowledge of the measurements of the space is needed.

There are a number of statistical measures that observe the transition from early reflections to late reverberation. Kurtosis is well-suited for determining the mixing time of a room as it indicates when a window of samples has a symmetric distribution. Higher order cumulants, such as kurtosis, contain amplitude and phase information unlike second order statistics (correlation) which are phase-blind. While moments describe deterministic signals, cumulants are measures for stochastic signals; if a non-Gaussian signal is mixed with a Gaussian signal, higher-order cumulants will ignore the Gaussian noise portion of the signal [Nikias and Petropulu, 1993]. Kurtosis then becomes an ideal measurement for determining when an IR transitions from a deterministic signal to a stochastic one.

The fourth order zero-lag cumulant of a zero-mean process is often referred to as kurtosis and can either be normalized or unnormalized. Here kurtosis will refer to the normalized definition.

$$\gamma_4 = \frac{\mathbf{E}(x-\mu)^4}{\sigma^4} - 3 \tag{4.2}$$

where $\mathbf{E}()$ is the expectation operator, μ is the mean, and σ^2 is the standard deviation.

To determine the mixing time, the kurtosis of the IR is calculated for a sliding 30 ms rectangular window. A window size of 20-30 ms is recommended by Abel and Huang [2006] and 25 ms by Rubak [2005] when statistically analyzing IRs as it corresponds to the integration time of the ear. During the early reflections, the kurtosis value tends remains positive, but once the IR has a symmetric distribution, the kurtosis value sharply decreases and remains around zero. The mixing time is defined as the most rapid decrease in kurtosis to occur within 200 ms after the direct sound. According to the characteristics of a set of IRs, the analysed region is adjusted in order to avoid erroneously early mixing times. This threshold, t_{thresh} is manually adjusted.

The mixing time is defined as:

$$t_{mix} = \text{argmin}\Big(_{t=} \gamma_4 (\sum_{k=t_{thres}}^{t_thresh+200} \omega(k)h(k))\frac{d}{dt}\Big) + 2 \tag{4.3}$$

The mixing time, t_{mix}, of the IR, $h(t)$ is the derivative of the kurtosis, $\gamma_4$and is relative to the direct. It does not include the delay τ; the windowing function $\omega(t)$ is a 30 ms rectangular window. The added value 2 is an arbitrary amount of time added to ensure that the early reflections are fully included and not prematurely truncated. The top pane in Figure 4.2 shows an IR with the delay, threshold, and the analysed portion indicated. The second pane shows the analysed portion of the IR and the mixing time with the third and fourth showing the kurtosis and derivative values with the mixing time.

This mixing time metric is novel, though does hold similarities to previous

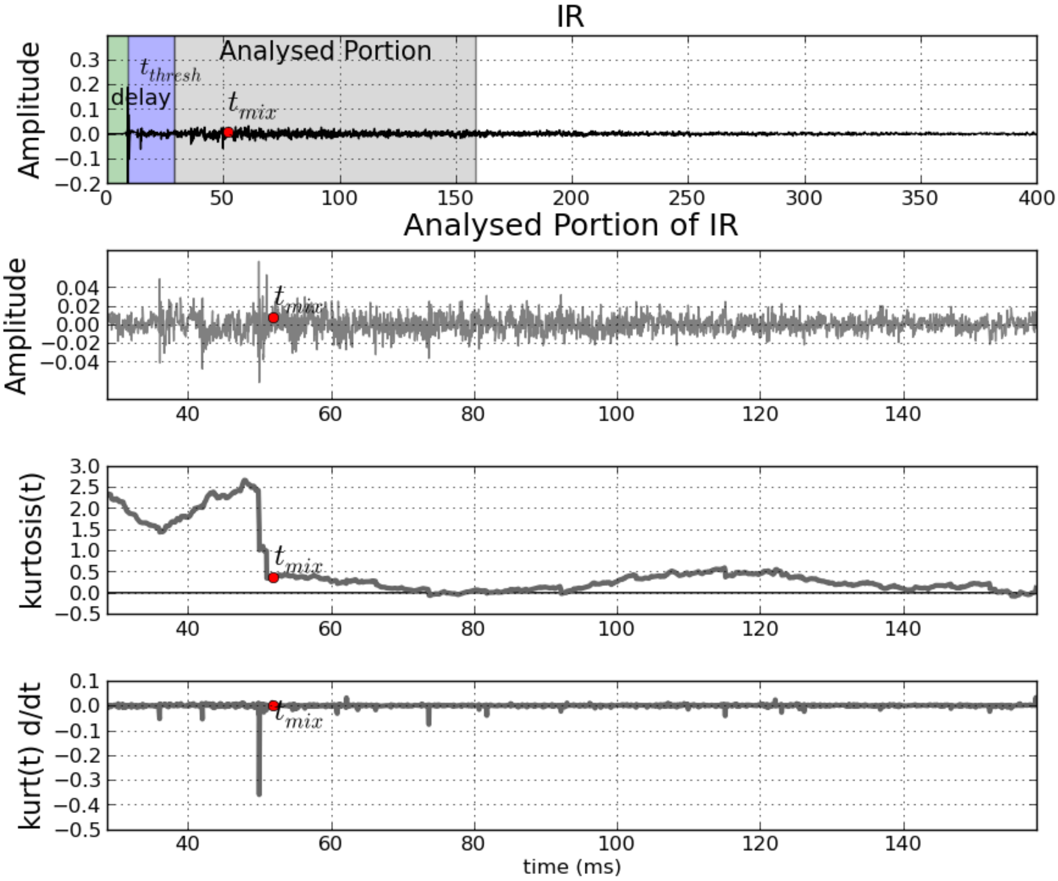


Figure 4.2: An IR: plotted in the time domain; the kurtosis; and the derivative of the kurtosis. The mixing time is indicated with a red circle.

approaches. Kurtosis was investigated as a measurement of echo density by Abel and Huang [2006], but was found to not be ideal. The kurtosis was determined for each window and normalised to the expected kurtosis for a Gaussian process. This was shown to have poorer performance when the echo density is small. By instead using the derivative of the kurtosis, changes in the kurtosis values are monitored instead of the absolute values.

While multiple IRs across a set have a similar mixing time relative to the rest of the set, there is some fluctuation. To allow for signal processing optimisation during synthesis, the set needs to be truncated to the same point in time, so a single mixing time needs to be selected. This done by examining the distribution of mixing times across the collection of IRs. A representative mixing time is then selected for the set. The mixing time is the longest time that best represents the set without including outliers. It is better to select a time that is too late than too early. A later time includes late reverberation from the room that could have been included in the late reverberation model; though less efficient, no information is lost. However, a shorter time can mean that information about early reflections is lost from the system. If the IRs have multiple channels, e.g. B-format, the process is repeated for each channel.

4.2.2 Panning and interpolation

To simulate a moving sound source with a set of HRTFs, a sound source is convolved with an HRTF pair and then cross-faded with the same sound source convolved with a spatially-adjacent HRTF pair [Begault, 2000, p. 132-136]. If the HRTF measurements are close enough together and cross-faded, then the sound source appears to move smoothly between positions. If the cross-fading rate is too slow comb filtering effects can occur while if the rate is too quick the switching between filters will be heard as artefacts such as clicks [Vorländer, 2008, p. 270].

Cross-fading between multiple measurements has been extended from HRTFs to IRs by Dalenbäck and Strömberg [2006]. CATT-Acoustic is an acoustic modelling program which calculates the IR of a measured space. To simulate real-time movement through the space it calculates a large number of full-length B-format IRs for a single source location and different receiver locations offline. Then in real-time it produces the auralization as defined by the user moving through the virtual space. This is done by interpolating between IRs and then performing a low-latency convolution and outputting the desired format of down-mixed B-format or binaural audio. Dalenbäck and Strömberg [2006] do not divulge how the IRs are interpolated other than the entire is used.

Reilly and McGrath [1995a,b], describe two similar systems for interactive binaural auralization. The system described in [Reilly and McGrath, 1995b] stored the coefficients for 128 binaural IRs – IRs with encoded HRTF spatial information. Using the Lake DSP Huron platform, they could convolve a signal in real-time with the appropriate IR in order to simulate head movements with head tracking. Though they could store up to 5 seconds of an IR sampled at 48 kHz, they found that dividing the IR into two filters, a 'head filter' containing the first 64000 samples (approximately 1.5 s at 44.1 kHz) and a 'tail filter' containing the remainder of the IR, was advantageous. The same 'tail filter' could be used for any point inside the room, while only the 'head filter' needed to be switched. In [Reilly and McGrath, 1995a] this system was extended to virtual Ambisonics. The HRTF information was no longer encoded with the room information. Instead, the early reflections were rendered in real-time according to the user's position and the late reverberation was generated by convolving the sound source with an IR measured in a real room. The computation time of the early reflections was decreased by using virtual Ambisonics.

A similar approach which combines these three systems, [Reilly and McGrath, 1995a,b, Dalenbäck and Strömberg, 2006] is investigated here. Multiple IRs are used to simulate a moving listener position, but the simulated effect is created by updating only the portion of the measured IR that is dependent on position rather than the entire IR as done in [Reilly and McGrath, 1995a,b] and not [Dalenbäck and Strömberg, 2006]. The system described in [Reilly and McGrath, 1995a] uses acoustic modelling software to calculate the early reflections a space. It does not use the full IR, but only the early reflections, similar to the system in [Reilly and McGrath, 1995b]. Like all three systems, virtual Ambisonics will be used here as well. While [Reilly and McGrath, 1995a] a single IR to generate the late reverberation, multiple IRs are used for late reverberation generation as will be discussed in Section 4.3.

If the measured set of HRTFs is not dense enough so that cross-fading between adjacent HRTF positions is not perceptually smooth, then interpolation is used to increase the number of HRTFs [Begault, 2000, p. 132–136]. Similarly, it is not necessary to measure the IR of all possible source positions, nor is it possible. However, the discrete measurements need to be close enough to one another so that a source does not 'jumps' from one position to another when a smooth transition is desired. Subsequent positions need to be located within the just noticeable difference or localisation blur which is dependent on the sound source type and location. See Table 2.1 for examples of localisation blur values.

Most previous studies of interpolating between acoustic measurements have involved HRTFs which don't contain room information. However, some recent work in [Ajdler, 2006] involves sampling and interpolating IRs from real and modeled spaces in order to find the IR for any arbitrary point in two dimensions. It was found that the interpolation errors greatly decrease when only the first 100 ms are interpolated. The mean squared error (MSE) was found to increase with time and it is postulated that this is because of the time-varying nature of temperature, humidity, and pressure within rooms. Only the early reflections will be interpolated here which are approximately 100 ms.

Interpolation of the early reflections of room IRs has been investigated most recently by Masterson et al. [2009] and previously by Huszty et al. [2008]. Masterson et al. [2009] were investigating interpolation within the context of wave-field synthesis to interpolate between receiver positions using dynamic time warping. Huszty et al. [2008] used a fuzzy model to interpolated different source positions.

The interpolation algorithm used here can be seen in Figure 4.3. Linear interpolation

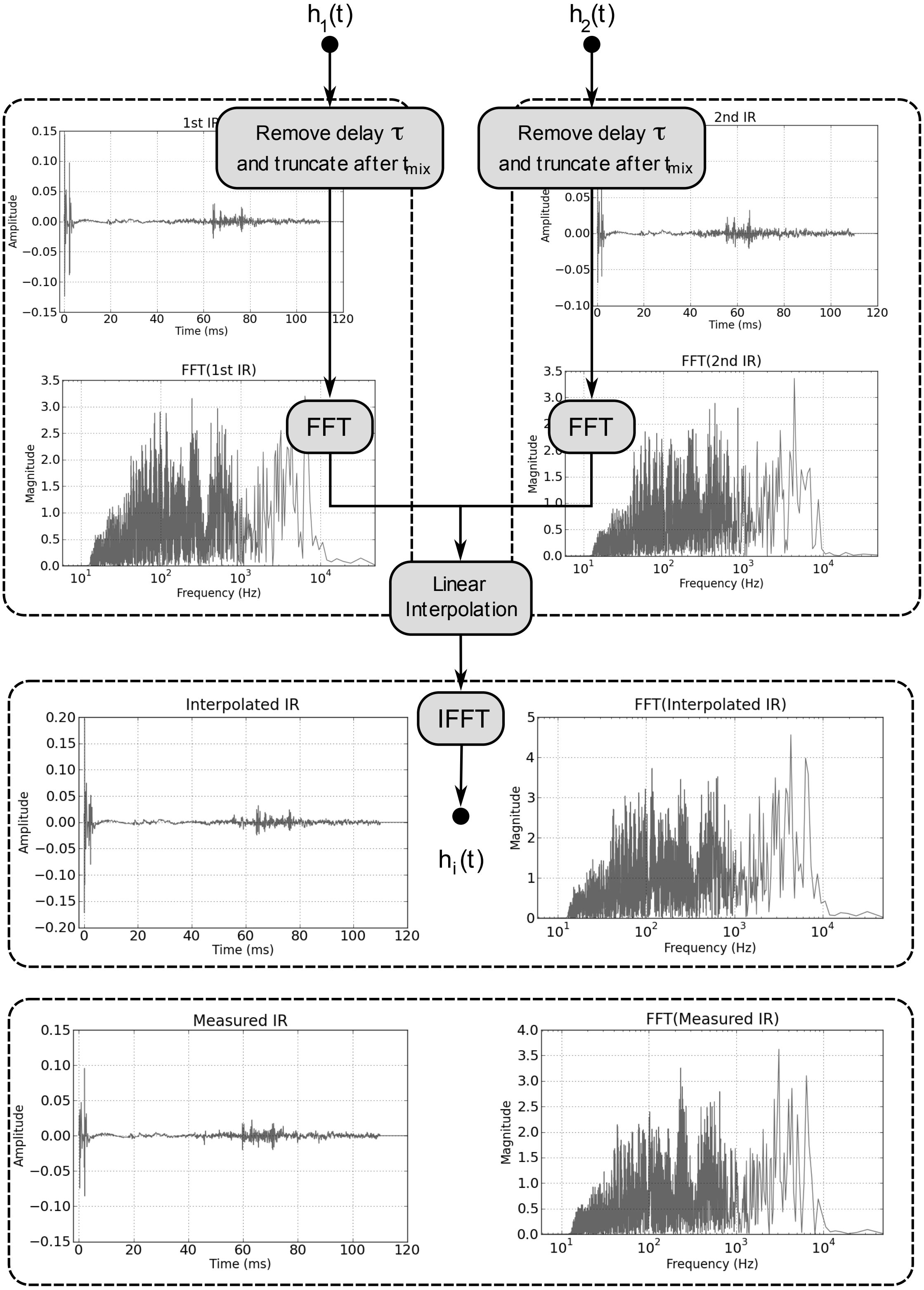


Figure 4.3: Algorithm for interpolating between IRs containing only early reflections.

is performed in the frequency domain for several reasons, primarily because it is the simplest method, especially since only two dimensions are currently being explored. However, Nishino et al. [1999] found that linear interpolation of HRFTs is sufficient when large, equidistant measurements are present; only small improvements are found with other interpolation methods. Additionally, it has been found that the most significant factor for decreasing error is computing the interpolation in the frequency rather than the time domain [Hartung et al., 1999]. It is acknowledged that linear interpolation does cause temporal smearing of reflections which is why [Huszty et al., 2008, Masterson et al., 2009] use more complex methods. However, linear interpolation of early reflections, and particularly B-format signals, has not been previously studied.

The interpolated measurement is created by interpolating between the IRs adjacent to the desired source location. The delay τ is removed and the IRs are transformed to the frequency domain. The magnitude and phase information are then linearly interpolated.

Given two IRs, h_1 and h_2, the interpolated IR h_i located on the line formed by the two IRs is:

$$h_i = \mathrm{IFFT}\left(\frac{\overline{h_1 h_i}}{\overline{h_1 h_2}}\,\mathrm{FFT}(h_1) + \frac{\overline{h_2 h_i}}{\overline{h_1 h_2}}\,\mathrm{FFT}(h_2)\right) \tag{4.4}$$

where FFT() and IFFT() are the fast Fourier transform and its inverse, and $\overline{h_1 h_i}$, $\overline{h_2 h_i}$, and $\overline{h_1 h_2}$ denotes the distances between the receiver positions of the IRs.

4.2.3 Evaluation

The data set used for the exploratory evaluation is a collection of IRs measured in three rooms at the Mile End Campus of Queen Mary, University of London: the Octagon, the Great Hall, and a classroom. The Octagon and Great Hall are significantly larger than the classroom. The Great Hall has a stage, floor seating and a balcony. As the name suggests, the Octagon has eight walls with a domed ceiling 23 m above the floor and is a regular octagon. Further details about the data set can be found in Appendix B.

4.2.3.1 Mixing time

Figure 4.4 shows the mixing times for the W channel of each receiver position for the Octagon. It is representative of the distribution of mixing times for the other channels of the Octagon and the Great Hall. The longest mixing times are the receiver positions closest to the sound source with the mixing times decreasing as the distance increases. The Great Hall and Octagon mixing times decrease to 50 to 60 ms in the farthest row of receivers. If the a mixing time cannot be calculated,it is found to be the earliest or latest time available. These appear as dark blue or red squares in Figure 4.4. A mixing time is more likely to not be found as the distance between the source and receiver increases and as the receiver approaches a boundary such as a wall or the balcony in the Great Hall.

The classroom is much smaller than the other two rooms and has a more consistent distribution of mixing times. There is not a significant change as the distance between the source and receiver increases.

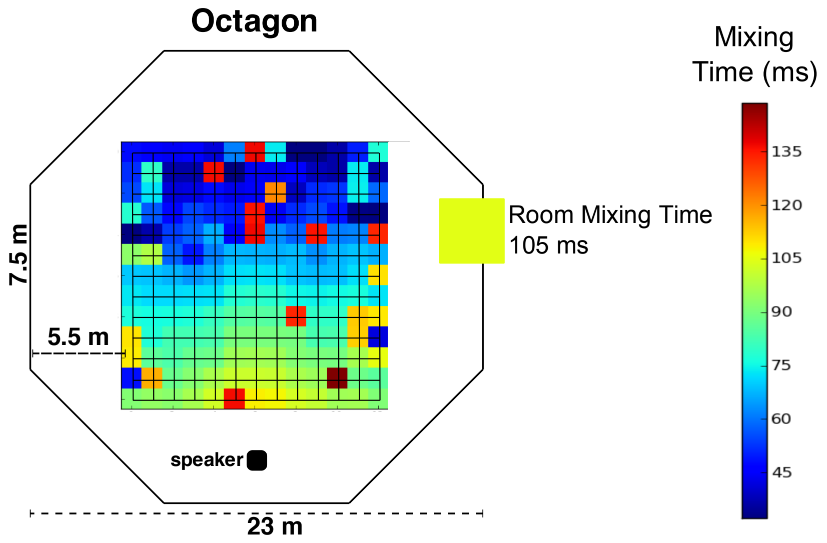


Figure 4.4: Distribution of t_{mix} for the W channel of the Octagon.

Room	B-Format Channel	Room t_{mix} (ms)	Percentage of IRs $\leq t_{mix}$
Classroom	W	49	75%
	X	54	81%
	Y	50	68%
	Z	55	81%
Great Hall	W	96	93%
	X	95	83%
	Y	100	91%
	Z	100	95%
Octagon	W	105	92%
	X	105	94%
	Y	94	77%
	Z	95	88%

Table 4.3: Mixing times for each Ambisonics channel for each of the three rooms.

The room mixing times for each B-format channel of each room are listed in Table 4.3. The mixing time for each channel is the maximum of the room and is an average of the closest receiver positions. The mixing times for all channels of the Great Hall and classroom are within 6 ms of each other. The Octagon has less consensus amongst its channels with the W and X channels having a mixing time 10 ms longer than the Y and Z channels. Both channels have a large number of receiver positions without a calculated mixing time that are near the sound source. The few positions that do have a mixing are around 95 ms, but it is not clear why this discrepancy occurs. The Octagon has a high domed ceiling approximately 23 m above the floor which may play a role in the earlier mixing times for the Z channel.

The right column of Table 4.3 shows the percentage of IRs whose determined mixing time is less than or equal to the ensemble mixing time. It is better for an individual IR to have a mixing time that is less than the ensemble mixing time than one that is greater. A shorter mixing time means the model is less efficient as more convolution is being performed than is necessary, but a longer mixing time implies that important early reflection information is being lost. The higher the percentage the more likely that all of the early reflections from every receiver position are represented.

4.2.3.2 Interpolation

To determine the best approach to interpolating early reflections, an IR is interpolated with different pairs of measured IRs and the mean square error (MSE) is compared. All interpolated IRs are located halfway between the measured pair. It is implied that an increase in MSE will result in greater audible artefacts or errors, but listening tests are required to determine the effect and severity of those errors.

The distance between measured IRs is increased to see how it influences the error. This is done for the same three rooms used in the above evaluation. Along with distance, the relative position of the two measured IRs in respect to the interpolated position and sound source is varied. The positions of the interpolated and measured IRs can be seen in Figure 4.5. Each room has a single interpolation position at the centre of the space. The measured IRs are then varied in four directions: directly in front and behind the interpolated position; to the left and right of the interpolated position; a diagonal to the left and front to the right and rear of the interpolated position; and a second diagonal to the front and right to the back and left of the interpolated position. For the Octagon

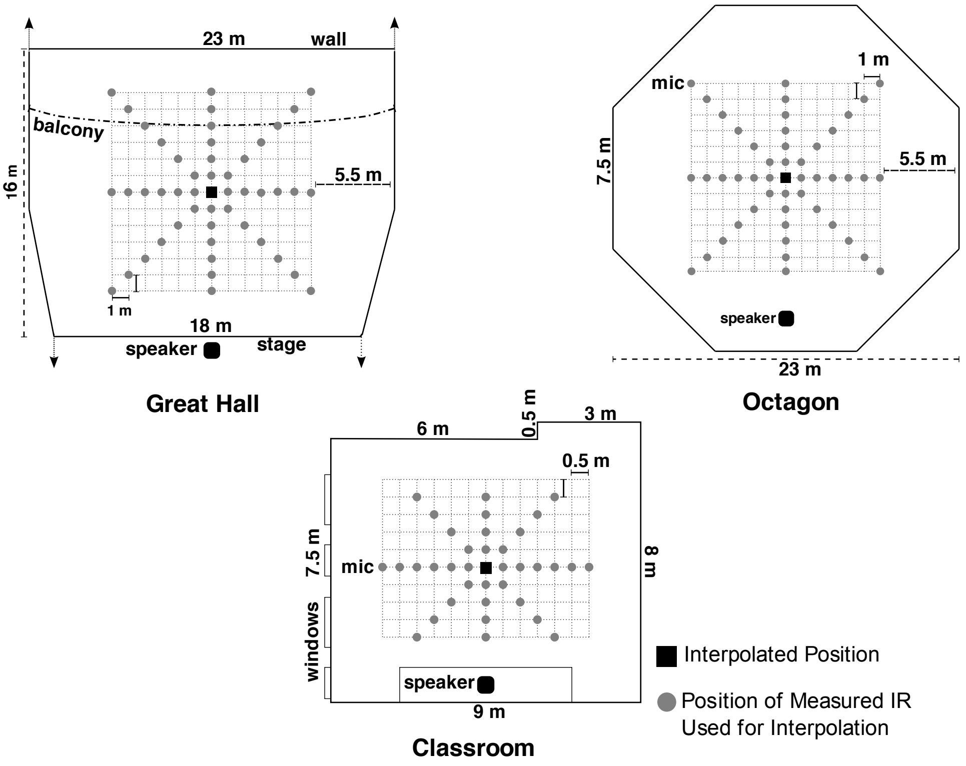


Figure 4.5: IR receiver positions used for interpolation. The square marker in each room indicates the interpolated position are the circles are the measured positions.

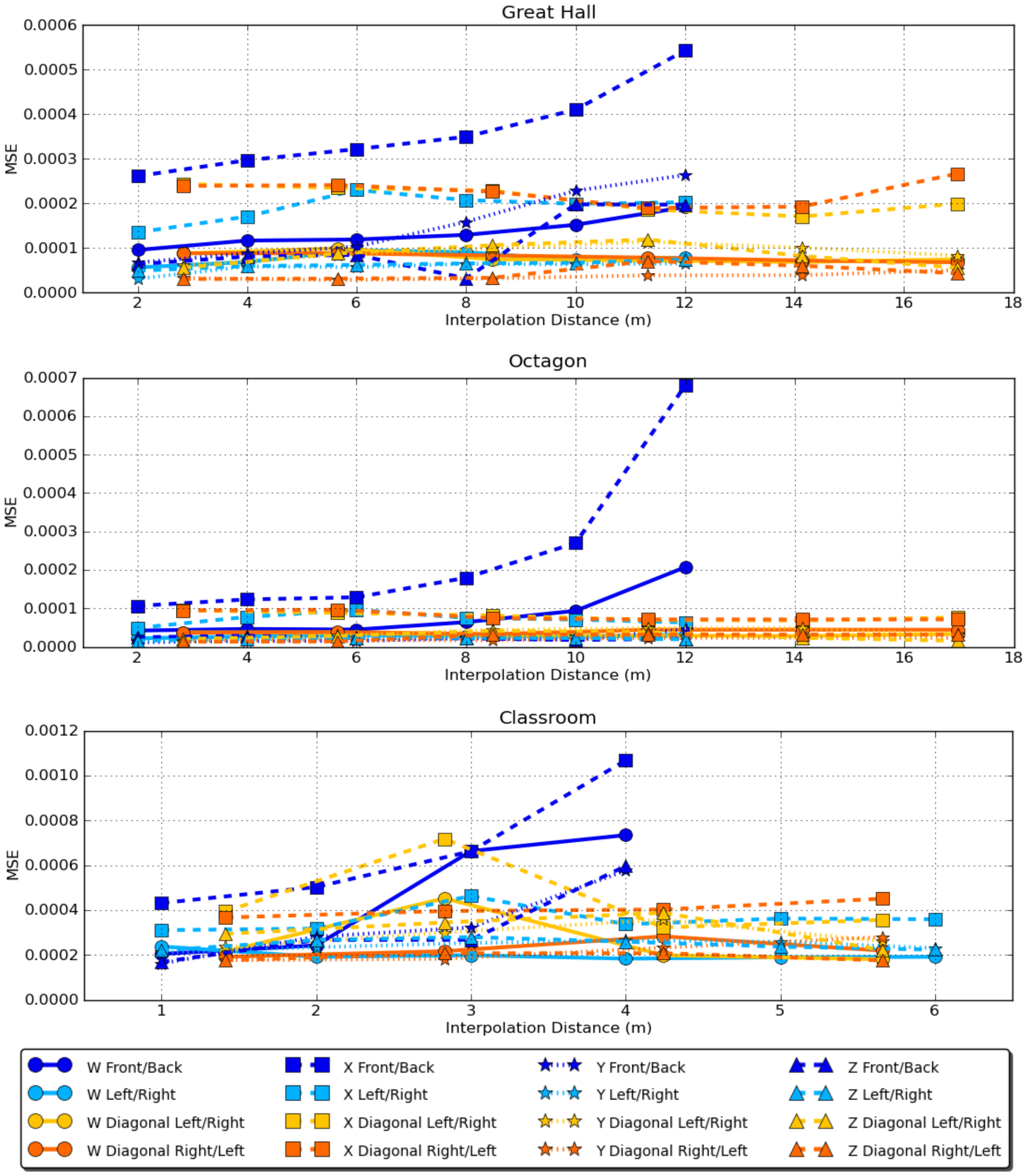


Figure 4.6: MSE versus distance for each B-format channel and direction in each room.

and Great Hall the front/back and left/right IR pairs are as close as 2 m and as far as 12 m increasing at intervals of 2 m. The diagonal pairs are 2.8 m to 17 m apart increasing at intervals of 2.8 m. The front/back and left/right pairs for the classroom range from 1 m to 6 m at intervals of 1 m. The diagonals range from 0.7 m to 5.7 m at intervals of 0.7 m.

The MSE for each B-format channel and direction of the IR pairs can be seen in Figure 4.6. While the differences cannot be easily distinguished between most of the channels and interpolation directions, the front/back direction for the majority of the channels and room stands out. It has significantly more error and increases that error faster as the distance between pairs increases. The front/back pairs in the Octagon show the largest variation. The rest of the channels and directions have very little change as the distance increases. In the Great Hall and classroom the X channels of the interpolation pairs that are at a diagonal to the interpolated position also show some more error than the remaining channels. The X channel of the front/back pairs for all three has the greatest error of the room.

While the MSE values may not directly translate to what artefacts in the audio may be heard, in general the greater the MSE, the greater the likelihood of artefacts. Overall, the MSE is related to the distance – as the distance increases, so does the MSE. When the measured IRs are the closest, the MSE is low. It is likely that interpolating from farther distances generates more audible errors.

4.2.4 Discussion

The evaluation of kurtosis as a means to determine mixing time consists of finding the mixing times of a collection of IRs measured across three rooms. The two larger rooms have longer mixing times around 100 ms with the smaller room with mixing time of approximately 50 ms. A single mixing time is found for each B-format channel in each room. Two of the rooms have a consistent mixing time across all of the channels, but the Octagon has a spread of 10 ms between the W and X and the Y and Z channels.

All three rooms have variations in the mixing time according to the receiver position. In general, for larger rooms the mixing time decreases as the distance between the source and receiver increases. In the smaller classroom, the mixing time is more consistent across all receiver positions. The variation in mixing time may be an artefact of using an empirical approach. When the data set used in the evaluation was measured, all IRs were measured with identical loudspeaker and microphone gains. The farther measurements therefore have a higher noise floor relative to the direct sound. Larger rooms then allow longer distances between source and receiver which can increase the number of errors.

The kurtosis method also has difficulties finding a valid mixing time as the receiver position approaches a physical boundary. This is seen in the classroom as the receiver approaches the back wall and in the Great Hall when the receiver approaches the back wall and balcony.

Lindau et al. [2010] have examined seven methods for determining mixing time, including a variation of the kurtosis method described here. (It does not find the derivative of the kurtosis values, but computes the mixing time as when $\gamma_4(t) = 0$.) They compared model-based approaches which calculated the mixing time based on parameters such as room volume and absorption empirical approaches which use signal processing of measured IRs. They compared how the mixing times determined by the seven approaches performed against listening tests. The mixing times of nine different shoebox shaped rooms were found. The results agree with what has been found here: the mixing time increases as the room volume and reverberation time increases. The mixing times of the Octagon, Great Hall, classroom are higher than those of the nine rooms with similar volumes, but they are not all shoebox shaped and it is difficult to make direct comparisons without a larger sample.

Lindau et al. [2010] also found high variance in the determined mixing time across all of the methods using an empirically measured IR when multiple IRs from the same room were compared. This variation also occurred when binaural room IRs were rotated (for head tracking). This implies, along with the findings here, that determining the mixing time from an empirically measured IR is typically not robust. Given that kurtosis is dependent on an IR being measured well above the noise floor and away from boundaries, it should not be used for calculating the mixing time of a single IR. However, when used to find an ensemble mixing time so that errors are distributed across the group, it can be used to determine a representative mixing time for the collection.

Evaluation of linear interpolation between early reflections of two different IRs found that the error is decreased when the closest pair of IRs to the position being interpolated that are both approximately equidistant from the sound source are chosen. The front/back interpolation pairs consistently have greater MSE than the other interpolation directions regardless of the B-format channel. The amount of error also increases with distance faster for the front/back direction than the others. The diagonal directions also have greater error than the left/right direction further indicating that it is best for the interpolation to be between points which are approximately the same distance from the sound source.

Listening tests are needed to determine the perceptual effects of linear interpolation, but it has been studied informally by Southern et al. [2009]. They interpolated between measurements 5.5 m apart and found that the time of arrival of the direct sound and early reflections is delayed in the interpolated B-format IR, but that the directional information has been preserved. It is not clear how the interpolation was performed other than linear interpolation was used. The data set used here has IRs only 0.5 m or 1 m apart, depending on the room, so the error could be expected to be much less than that encountered by Southern et al. [2009].

It has been found by Blauert [1997] and Ajdler et al. [2005] that in a non-reverberant environment, sounds can be localised in the azimuth plane to as few as 5°, which increases when sounds move off-axis. This is reflected in HRTF libraries by Gardner and Martain [1994] and Algazi et al. [2001]. However, there are relatively few extensive studies of localisation in the azimuth plane with full room information; many more choose to control the environment and use anechoic or semi-anechoic environments [Begault, 2000]. However, it has been found that the ability to correctly localise a source can be greatly reduced in reverberant environments [Begault, 1992]. Room geometry, the type of source sound material, and the reverberation time of the room all affect the localisation and early lateral reflections, which are desired in musical spaces to give a broad image, can interfere with localisation [Hartmann, 1983].

In Hartmann [1983], subjects were able to correctly localise a sound source within 5° in the azimuth plane for a variety of conditions. The distance between the source and receivers in the three rooms examined ranges from approximately 2 m to 20 m. A change of about 5° then translates to about 0.2 m to just under 2 m. So at farther source-receiver distances, further IRs may not need to be interpolated as a change in receiver position of 0.5 m or 1 m may not cause any change in perceived location.

However, closer source-receiver positions would probably need to be interpolated for smooth movement. For the best results, the two closest IRs that are approximately equidistant from the source should be used for interpolation.

The questions that remain then are: how close together do IRs need to be, i.e. what is the JND in quantisation before interpolation creates too many errors; what are the best cross-fading parameters between early reflections; and how does kurtosis compare with other mixing time measurements.

4.3 Spatial Averaging of Late Reverberation

Previous hybrid reverberation research has used recursive statistical reverberators such as FDNs to model the late reverberation [Browne, 2001, Radford, 2003, Stewart and Murphy, 2007, Greenblatt et al., 2010]. Hybrid reverberation was initially developed to decrease the computational burden of convolution reverberation whilst retaining the accuracy of the IR. As processors can easily accommodate one channel of real-time convolution, networks of delay lines now have diminished advantages over convolution as artificial reverberators. Recursive delay networks are difficult to tune to precisely achieve a specific time-frequency response, but white noise can be shaped to mimic a specific response and then used in convolution reverberation. Recursive delay networks were used as a single representative late reverberation tail, and this can still be done with convolution. A time-frequency envelope representative of an entire room just needs to be extracted and used to shape white noise.

To review Section 2.2, the portion of the IR past the mixing time, the late reverberation, can be assumed to be a stochastic signal. Unlike the direct sound and early reflections, its temporal characteristics are independent of receiver position with the relative strength of modes being the greatest change [Jot et al., 1997]. As reverberation tails are assumed to be noise-like, then averaging sample by sample across a collection of IRs is destructive and will not produce a representative IR. The defining characteristic of a reverberation tail is the frequency decay over time, so the time-frequency envelope of each IR in a set is extracted and averaged. The average envelope can then be applied to white noise.

The average time-frequency envelope is analysed and then the representative late reverberation tail is synthesised in the following steps:

1. Calculate the time-frequency envelope for an IR after the mixing time

2. Repeat for all IRs in the set

3. Find the mean of the envelopes

4. Generate white noise

5. Convolve the average envelope with the white noise

4.3.1 Analysis

An Energy Decay Relief (EDR) is a Schroeder Energy Decay Curve (EDC) separated into frequency bands [Jot et al., 1997]. An EDC is the backwards integration of a squared IR over time and is usually used to compute reverberation time (see Section 2.2.5). EDRs were developed as a method to remove noise from an IR by extracting the envelope and calculating the noise floor. The portion of the EDR representing the noise floor is removed and the EDR of the IR is extrapolated past the noise floor. The EDR is then convolved with white noise to synthesis the IR without any measurement noise. Here multiple EDRs are averaged to calculate a representative time-frequency envelope of a collection of IRs.

To calculate the EDC of an IR:

$$EDC(t) = \int_{t}^{+\infty} h^2(\tau) d\tau \tag{4.5}$$

where $h(t)$ is the impulse response at time t. This means that $EDC(t)$ is equal to the remaining energy in the IR after time t. The extension from EDC to EDR is discussed in more detail by Jot et al. [1997], but we will summarise the definition of an EDR in Eq. 4.6 and 4.7.

Let $h(t)$ be a time domain IR and ρ a time-frequency distribution function.

$$h(t) \xrightarrow{\rho} \rho_h(\tau, f) \tag{4.6}$$

$$EDR(t, f) = \int_{t}^{+\infty} \rho_h(\tau, f) d\tau \tag{4.7}$$

When the short-time Fourier transform (STFT) is the time-frequency representation, the EDR is the backwards integration of the squared magnitude of each bin of the STFT over time.

Measurement noise can inflate reverberation time calculations, particularly in the high frequencies. To help prevent this, the magnitude of the frequency content of the last window of the IR is subtracted from the magnitude of the frequency of all preceding windows before the EDR is calculated. It is acknowledged that this is a naïve approach and that the technique described by Jot et al. [1997] could be used in the future.

As the stochastic model of reverberation is valid only for the diffuse portion of an IR, the IRs are only analysed after the mixing time determined by the kurtosis metric described in Section 4.2.

4.3.2 Synthesis

The synthesis of the averaged tail is two steps. White noise is generated and windowed with a rectangular window the same length and hop size as the analysis windows. Each window of the white noise is multiplied with the corresponding window from the averaged EDR, convolving the envelope with the noise. The windows are then transformed back into the time domain and overlap-added together.

4.3.3 Evaluation

The analysis-synthesis technique to generate an averaged late reverberation tail is compared against each of the measured IRs from the three acoustic spaces described in

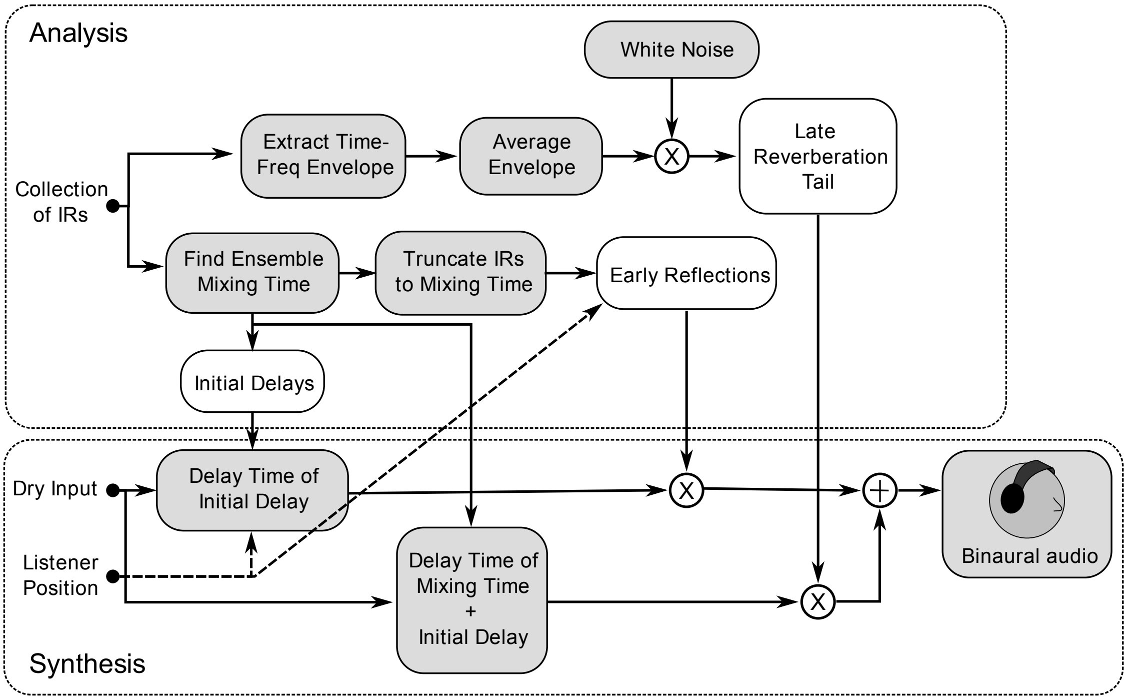


Figure 4.7: Signal flow for full interactive room model.

Appendix B. Common metrics used for classifying acoustics spaces are used to observe how the synthesised IR differs from the original IR.

An IR is synthesised by passing an impulse through the system. The impluse is delayed according to the initial time delay, then convolved with the early reflections and the averaged late reverberation as shown in Figure 4.7. A real-time implementation of the system would include updating the initial time delay and early reflections according to the desired receiver position, but only static measurements are used in this evaluation. For evaluation, the tail is combined with the appropriate set of early reflections so that when compared with the measured IR, the only difference in the signal is the late reverberation tail.

Two standard acoustics metrics, clarity and reverberation time, are selected from ISO 3382 [1997] to evaluate whether the resynthesis of the IRs introduces errors when compared to the measured IRs. Both reverberation time and clarity have been shown to be perceptually-relevant acoustics metrics [Beranek, 2004]. Since synthesising the tail may alter the ratio of early to late energy in the IR, clarity is calculated for both speech (C_{50}) and music (C_{80}) applications. The reverberation time (RT_{30}) across eight octave bands is measured to compare the frequency-dependent decay rate.

4.3.3.1 Clarity

Clarity measures the ratio of early to late energy with the distinction between early and late being 50 ms for C_{50} and 80 ms for C_{80}. Section 2.2.5 discusses clarity measures in more detail.

Table 4.4 lists the average C_{50} and C_{80} error for each room. There is very little error incurred with mean errors of 0.06 dB or less for the majority of the channels and

		C_{50} (dB)		C_{80} (dB)	
Room	*Channel*	*Mean Error*	σ	*Mean Error*	σ
Classroom	W	0.00	0.01	-1.25	0.18
	X	0.00	0.02	-1.57	0.27
	Y	-0.36	0.07	-1.58	0.17
	Z	-0.01	0.03	-1.34	0.15
Octagon	W	0.00	1.35	0.00	1.32
	X	0.00	1.77	-0.01	1.67
	Y	-0.04	1.71	-0.06	1.21
	Z	-0.04	1.97	-0.50	1.36
Great Hall	W	-0.03	0.02	-0.04	0.03
	X	-0.02	1.96	-0.02	0.04
	Y	-0.03	2.91	-0.04	0.03
	Z	-0.03	2.89	-0.05	0.03

Table 4.4: The mean difference and standard deviation in clarity measures between corresponding receiver positions of the synthesised and measured IRs.

rooms. However, for all three rooms the average clarity is lower for the synthesised IRs than the measured IRs. As all three rooms have mixing times around 50 ms or 100 ms, when the C_{50} values of the synthesised and measured IRs are compared, the early energy remains the same with only the late energy changing. The decreased C_{50} values for the synthesised IRs indicates that the spatial average tail has more energy than the original tails. The C_{80} errors for the classroom are greater, but the mixing time is around 50 ms, so the 80 ms early portion of C_{80} includes a portion of the averaged late reverberation tail along with the early reflections.

The errors are dependent on receiver position for the two larger rooms, but the errors are more evenly dispersed in the smaller room. The C_{80} error of the Great Hall can be seen in Figure 4.8; the Octagon has a similar pattern. In the two larger rooms, the farther the receiver position is from the sound source, the more likely the clarity is to be too low. The closer positions to the source better approximate the clarity of the measured IRs, though the clarity is too high for several of the nearest positions. The averaged tail is a better representation of the nearer receiver positions, at least in regards to clarity. The exception is the Z channel of the Great Hall which has very little error for the entire room except underneath the balcony directly in front of the sound source. The lower ceiling boundary of the balcony causes relatively large errors.

The smallest room, the classroom, does not have the position dependent error seen in the Great Hall and Octagon. Instead, there are large errors spread randomly around the room.

4.3.3.2 Reverberation Times

Frequency-dependent reverberation time ($RT_{30,f}$) is a measure of the time-frequency envelope of reverberation. RT_{30} is the time an IR takes to decay 60 dB and is extrapolated from the time the sound takes to decay 30 dB. See Section 2.2.5 for a review of how

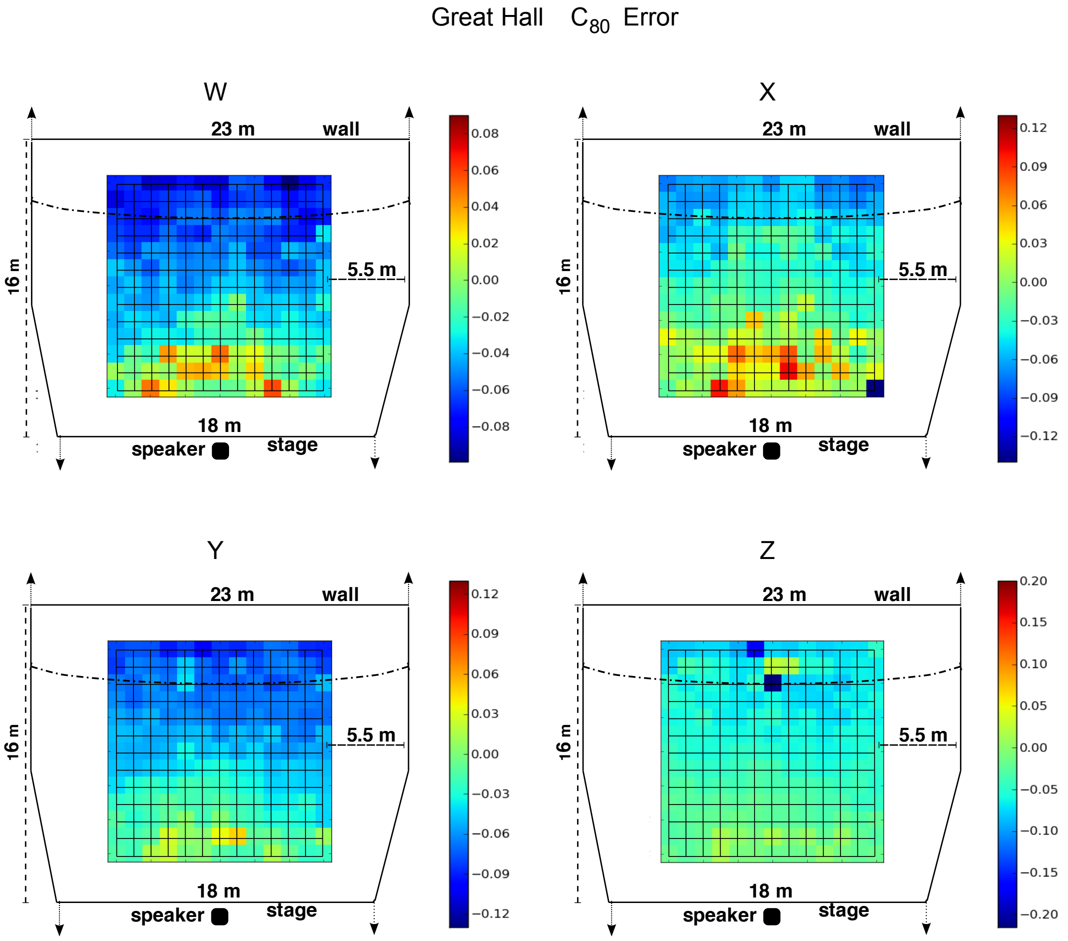


Figure 4.8: Difference in C_{80} for all receiver positions for the synthesised and measured IRs from the Great Hall.

reverberation time is calculated. For evaluation, the early reflections for each measured IR are combined with the average tail to create the synthesised IRs. The reverberation time of the octave bands from 125 Hz to 16 kHz for the measured and synthesised IRs are compared.

At higher frequencies the synthesised IRs have RT_{30} values similar to the measured IRs. For all three rooms, the mean reverberation times are within 0.3 s of the measured IRs above 500 Hz. Figure 4.9 show the mean difference in reverberation time for each frequency band at all receiver positions. The error bars show the standard deviation from the mean.

The Great Hall differs from the other two rooms as it does not have greater error in the lower frequency bands. Also, up to 4 kHz, the X channel has longer reverberation times than the measured IRs while the rest of the channels have shorter. The other two rooms, the Octagon and classroom, have large variations in error in the low frequencies with shorter reverberation times at 125 Hz and longer reverberation times than the measured IRs at 250 Hz. The synthesised IRs of the Octagon and classroom also have consistently lower $RT_{30,125}$ values, regardless of receiver position. At 250 Hz, the reverberation times are consistently too high across both rooms, again with no relationship to receiver position.

Figure 4.10 shows the error between the measured and synthesised IRs for each receiver position for each channel at 1 kHz. The receiver positions are in the same configuration as the receiver positions in Figure 4.8. The Octagon and classroom are similar while the Great Hall differs greatly. The reverberation time error in the Octagon and classroom are not highly dependent on receiver position, unlike the Great Hall which has increased error towards the sound source. Also, the reverberation times of the Octagon and classroom are slightly shorter than the measured IRs while the Great Hall has reverberation times that are slightly too long.

4.3.4 Discussion

The averaged late reverberation tail synthesised from the average EDR of a collection of IRs performs well under certain conditions, but is not fully representative of every receiver position within a room. The errors from the averaged tail are evenly distributed across the smallest room, the classroom. This is seen with both the clarity and the frequency-dependent reverberation times. As it is the only room of its size evaluated, it is not clear if these results are a function of the small room, the spacing of the receiver positions, or a combination of both. The two larger rooms have a receiver spacing of 1 m while the classroom has measurements every 0.5 m.

The two larger rooms have errors dependent on the location of the receiver. In general, the receivers nearer the sound source have less error in the clarity and more error in the reverberation time, but the Octagon and Great Hall perform differently from each other. The reverberation time error of the Great Hall is more dependent on receiver position than the Octagon.

The mean synthesised clarity (C_{50} and C_{80}) values are lower than the measured

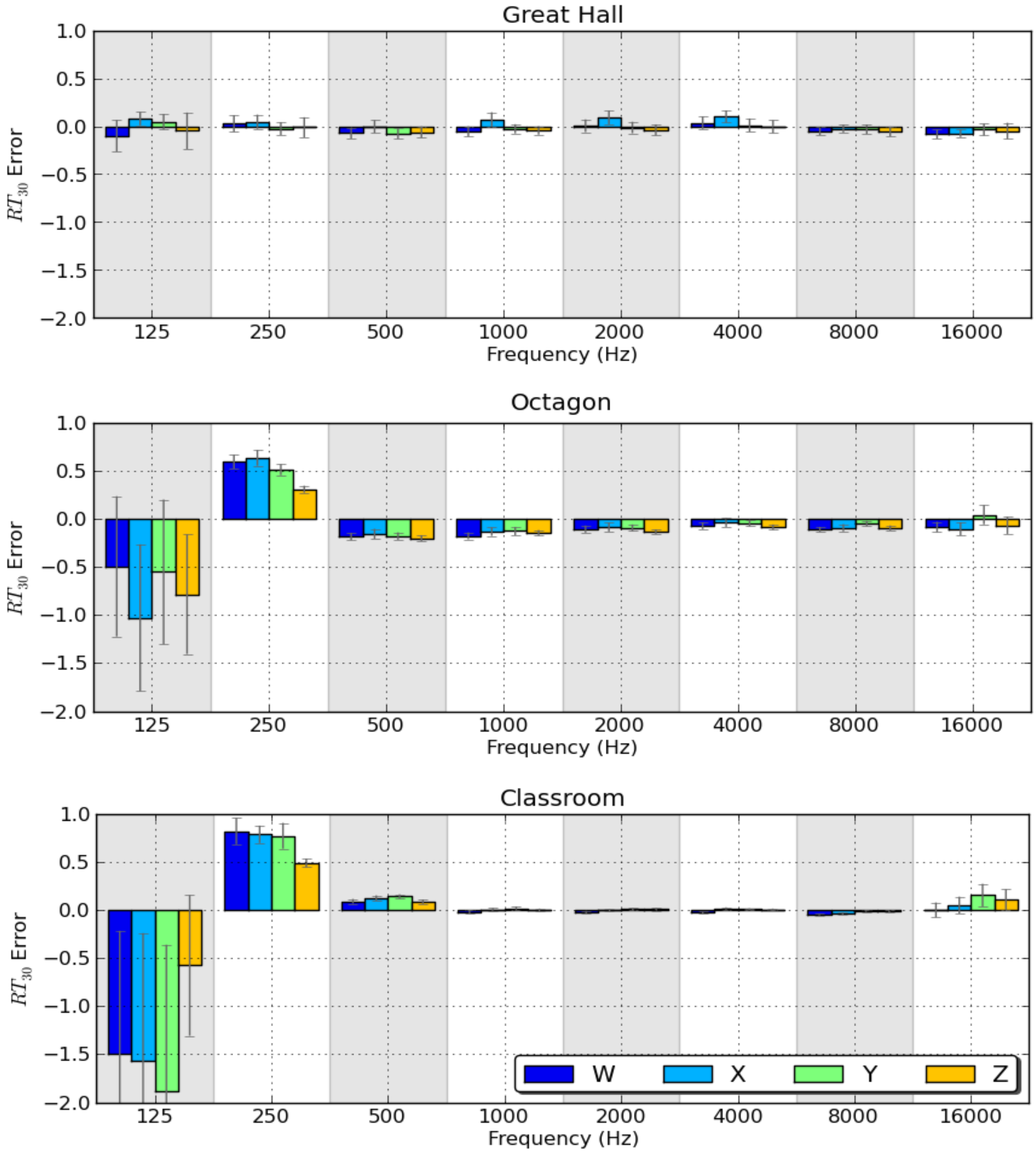


Figure 4.9: Mean difference in reverberation times from 125 Hz to 16 kHz across all receiver positions between the synthesised and measured IRs. The error bars show the standard deviation. Positive values indicate where the RT_{30} of the synthesised IRs are longer than the measured.

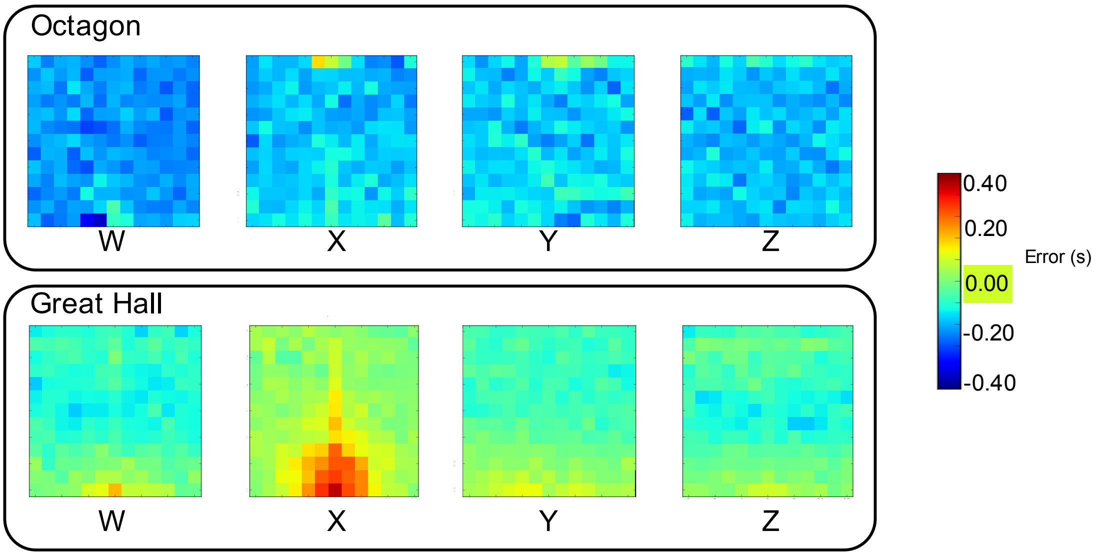


Figure 4.10: Mean difference in $RT_{30,1000}$ across all receiver positions between the synthesised and measured IRs. The error bars show the standard deviation. Positive values indicate where the $RT_{30,1000}$ of the synthesised IRs was longer than the measured.

IRs, but frequency-dependent reverberation time error of the Great Hall differs from the Octagon and classroom. The Octagon and classroom have similar errors in lower frequencies with short $RT_{30,125}$ times and long $RT_{30,250}$ times. The lower frequencies also have large variations in the error between receiver positions. The variation in the frequency response is likely to be a symptom of room modes, but it is not clear why the Great Hall does not exhibit the same variation as the Octagon and classroom.

The analysis-synthesis technique provides satisfactory results, but still leaves areas requiring improvement. On average, the synthesised IRs achieved C_{50} values within 0.60 dB and RT_{30} values within 0.3 s for most frequency bands of the measured IRs.

4.4 Summary

Complex acoustics models can create realistic auralizations of historical spaces, but these models are not interactive. Chapter 4 presented a new reverberation model based on hybrid reverberation which used a collection of B-format IRs to simulate real-time movement. A new metric for determining the mixing time of a room was developed and interpolation between early reflections was investigated. Though hybrid reverberation typically uses a recursive filter such as a FDN for the late reverberation, an average late reverberation tail was instead synthesised for convolution reverberation.

Hybrid reverberation requires the separation of early reflections and late reverberation from an IR. The mixing time is the earliest point in time that an IR can be considered stochastic and the room mixed. A number of statistical measures can be used to observe the transition of the deterministic early reflections into stochastic late reverberation. Kurtosis was used here. This approach is advantageous as it does not require any knowledge about the room such as volume or absorption coefficients. Other metrics that do not require knowledge about the space such as Abel and Huang [2006] are used to indicate how quickly echo density builds. They are not explicitly designed to determine the earliest point in time an IR can be represented by a stochastic model. Evaluation of the mixing time metric found that it is useful for determining the mixing time of multiple IRs from the same room. However, it is not robust enough to determine the mixing time of an arbitrary IR.

Linear interpolation can be used to increase the sampling density of a room. Evaluation of linear interpolation between early reflections found that the error is decreased when the pair of IRs used in the interpolation are the closest pair of IRs to the position being interpolated and are both approximately equidistant from the sound source are chosen.

A technique to find a representative late reverberation tail for a set of IRs measured in the same space has been described. The diffuse portion of each IR is analysed and the time-frequency envelope extracted. The envelopes across a set of IRs measured in the same room are averaged and then convolved with white noise. During real-time synthesis, a dry sound source is convolved with the shaped white noise. This decreases the computational complexity and storage requirements of real-time convolution as the same coefficients are used for any receiver position. The tail is combined with early reflections and delayed according to the receiver position.

To evaluate how representative the averaged tail is for a room, three room were modelled. The average tail was added to the early reflections of each IR to produce the synthesised IRs. Two standard acoustics metrics, reverberation time and clarity, were used to evaluate the differences between the measured and synthesised IRs. Results were not consistent across all three rooms. The Great Hall produced better frequency-dependent reverberation times than the other two rooms, though all three rooms can closely simulate the reverberation time on average to within 0.3 ms at higher frequencies. The classroom and Octagon had much greater error at lower frequencies.

The clarity, both C_{50} and C_{80}, was more consistent across all three rooms. In each room, the clarity of the synthesised IRs was lower than the measured IRs indicating that the analysis-synthesis technique produces a reverberation tail with more energy than the measured IRs.

Determining the mixing time of a room and interpolating between IRs have been investigated in the literature, but creating a representative reverberation tail has not. It is not possible to compare this technique to the other approaches. While the model does not perfectly replicate every receiver position within a room, it can closely approximate most positions. The model may be acceptable for archaeological applications. Listening tests are needed to qualify the differences between the measured and modelled rooms.

Chapter 5

Music Collection Auditory Displays

Music is consumed by listening, yet listening is seldom an integral part of music browsing or search systems. When audio playback is integrated into a browsing or search system, users are able to find content much more efficiently and with greater satisfaction [Arons, 1997, Fernström and Brazil, 2001, Hamanaka and Lee, 2006, Heise et al., 2008, Ali and Aarabi, 2008]. Furthermore, it has been shown that when users listen to music rather than only read the associated metadata, they make different decisions [Barrington et al., 2009].

To review Section 3.1, auditory displays convey information via audio as opposed to visual displays which use text and graphics. In graphical user interfaces, users expect to be able to directly interact with icons and graphics and have immediate feedback from their actions, known as direct manipulation. Direct manipulation with audio content is less common, but is also important especially when the task is to retrieve or discover audio content. Auditory display which allows for direct manipulation of the audio data is called direct sonification. With direct sonification "the information sought is also the information displayed" [Fernström, 2005]. If the information sought is a specific melody from a song, then the information displayed is the playback of that song, not only the metadata describing that song.

Music information retrieval (MIR) applications that involve seeking or browsing audio content may be intended for an expert user with a highly specific task or for a general consumer engaging in serendipitous browsing. A musicologist may need to skim the repertoire of a musician to gain insight into the performances, or a student may be looking online for a song just heard on the radio. In both cases searching and browsing are employed and in both the end user is a human. Therefore human factors do need to be considered for music browsing or search interfaces.

When audio is integrated into a search or exploration interface, playback typically only occurs serially; one song is played at a time and only if explicitly instructed to do so by the user. This can mean that poor metadata or lack of familiarity can inhibit a search or browsing task. After observing how the public browses and searches for music in music stores and public libraries, Cunningham et al. [2003] conclude that novel software systems for music discovery may need to create new interaction paradigms. They mention Sonic Browser [Fernström and Brazil, 2001] as an example of a novel means to interact with a collection of music. It is the first spatial auditory display of its kind, and as will be

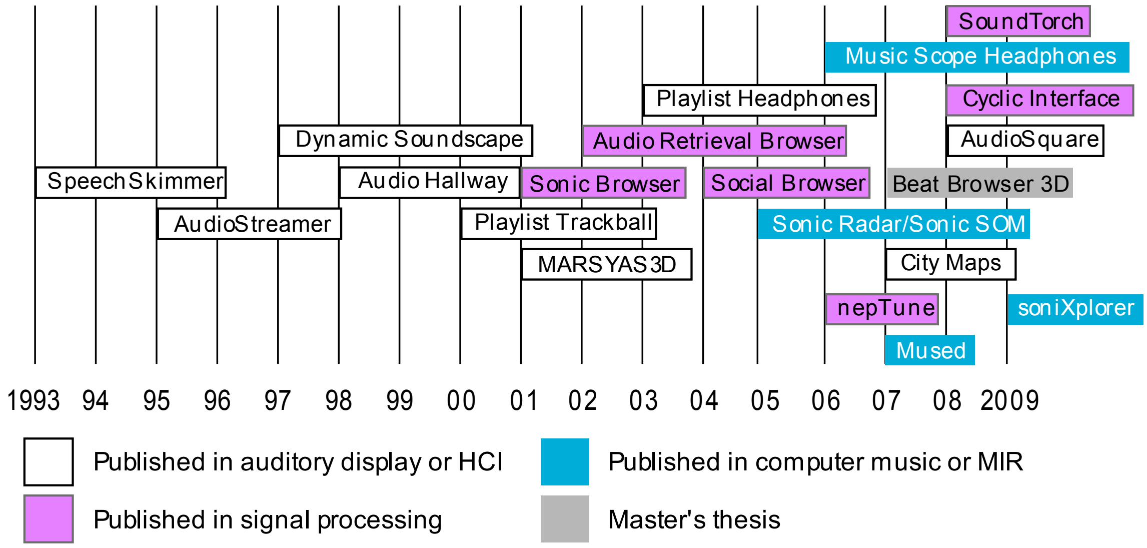


Figure 5.1: A timeline of when each interface was first published and the community where it was first published.

discussed, a design continually revisited. Sonic Browser and similar interfaces use passive presentation of audio content. Users do not need to make a litany of decisions before they listen to any audio; the audio is presented first and then further information such as metatdata can be learned if so desired.

The cocktail party effect, the ability to shift focus and attention amongst a number of audio sources playing concurrently, is advantageous in audio user interfaces. In particular, MIR interfaces which are trying to convey information about a selection of audio files can use the cocktail party effect to present audio in a more efficient manner than serial presentation (playing a single audio file at a time). Auditory display is the presentation of information via audio while spatial auditory display explicitly uses spatial audio in its presentation.

This chapter will establish the common design features found in a wide number of music browsing and search interfaces with spatial auditory display. A brief overview of interfaces employing spatial audio as a significant component of interaction will follow. Not all of the discussed interfaces use spatial audio, but all do contribute to how auditory display can enhance search and browsing tasks of audio content. The common design features are compared and conclusions are drawn with suggestions as to how new interfaces can build upon past successes.

5.1 Overview of Previous Interfaces

This chapter reviews 20 interfaces with publication dates ranging from 1993 to 2009 published in the auditory display, music information retrieval (MIR), audio signal processing, and human-computer interface (HCI) communities. Figure 5.1 illustrates where and when these interfaces were initially published, and Figure 5.2 illustrates the key implementation details for each interface. All of the interfaces use audio as a primary means of interaction though some use visuals as well. Not all of the interfaces are

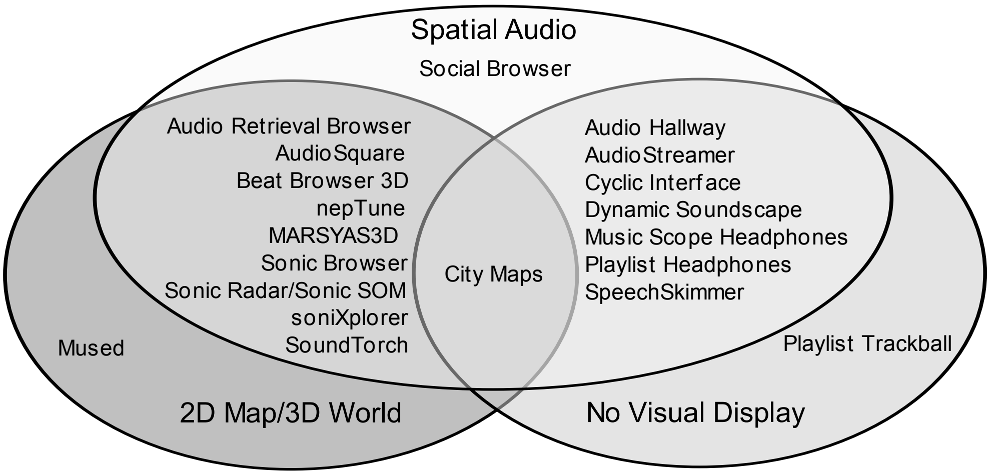


Figure 5.2: An overview of all the interfaces examined and the distribution of interfaces: using spatial audio; that are built around a two-dimensional or three-dimensional mapping of an audio collection; and those that do not require a visual display.

explicitly designed for music search or browsing tasks, but could be adapted to do so. Most of the interfaces use spatial audio and those that do not were selected because they contribute to how auditory display can be used to interact with a collection of music.

5.1.1 Earliest interfaces

Some of the earliest work exploring audio as a primary means to interact with a collection of speech or music comes from the Massachusetts Institute of Technology (MIT) Media Lab. SpeechSkimmer [Arons, 1993, 1997] navigates speech signals using a variety of approaches to increase the speed of presentation without impinging on intelligibility. AudioStreamer [Schmandt and Mullins, 1995] builds on SpeechSkimmer by presenting a user with three concurrently played speech signals placed in static positions, directly in front of and 60 degrees to either side of the user.

AudioStreamer developed into the Audio Hallway [Schmandt, 1998], a virtual acoustic environment modeling a hallway lined on either side with a series of doors. The user hears summaries of the collection of audio contained in the adjacent virtual rooms as they pass each virtual door. Each room contains speech content related to a common topic. When the user chooses to hear more about the summarised content behind a door, the user enters the room. The audio content within that room is then spatially arranged around the user's head.

Further work from the MIT Media Lab produced Dynamic Soundscape [Kobayashi and Schmandt, 1997] which assists in quickly finding specific portions of an audio file without needing to listen to the entire file serially. The main component of Dynamic Soundscape is the *speaker*, a sound source constantly rotating around the user's head. The intent is to associate the content of a speech signal with a spatial location. There can be multiple *speakers* at the same time at different locations around the user and each *speaker* may be playing audio from the same file but each starting at different point in time within the file.

Hiipakka and Lorho [2003] describe an interface which we will refer to as Playlist Headphones. The interface does not require any visual information, but lets a user navigate a genre-artist-album-song hierarchy to select songs to build a playlist. Menu options are spatially mapped from left to right with text-to-speech identifying textual metadata. A similar interface without any spatial audio elements is implemented by Pauws et al. [2000], an interface we will call Playlist Trackball. A force feedback trackball and an optional visual display is used to navigate a similar menu as used in Playlist Headphones, but also includes a music recommender. Music Scope Headphones [Hamanaka and Lee, 2006] is another interface using audio information without any visual information to select a song from a collection. Music Scope Headphones is an audio-only interface meant to assist users in choosing a single song from a selection of ten songs or listening to a multi-track recording and interacting with individual tracks.

5.1.2 Orientation perspectives

The *orientation perspective* or *museum metaphor* described by Cohen [1991] has static sound sources and a moving listener. The following interfaces use this perspective to navigate a collection of music, though the audio cursor described by Cohen as part of the *egocentric perspective* is also employed in many of these interfaces.

While audio-only interfaces are particularly beneficial for applications which have reduced or no visual displays such as mobile devices, visual information can compliment auditory display. The first interface using audio and visual information for MIR tasks is the Sonic Browser [Fernström and McNamara, 1998, Brazil and Fernström, 2003, Fernström and Brazil, 2001, Fernström, 2005]. The initial interface arranges audio files in a two-dimensional space along with a visualisation which also maps other parameters such as file size to the images representing the files. A user then controls a circular cursor called an *aura* where any file icon contained within the aura would be played concurrently in a stereo space. The user can control the size of the aura and its placement in the two-dimensional map. Later versions incorporate additional visualisations of audio collections including hierarchical information [Brazil and Fernström, 2003]. Sonic Browser has also been combined with a powerful software tool, MARSYAS [Tzanetakis and Cook, 2000] to create the Audio Retrieval Browser [Brazil et al., 2002]. The Audio Retrieval Browser is very similar to Sonic Browser, but uses the powerful audio signal processing tools available in MARSYAS to automatically create two-dimensional arrangements of audio files. MARSYAS also has some spatial auditory display capabilities called MARSYAS3D [Tzanetakis and Cook, 2001]. With MARSYAS3D individual sound files are mapped to as many as 16 loudspeakers, and audio is played concurrently from each speaker.

Exploring music mapped to a two-dimensional space is a popular approach and one that is further explored in many other interfaces. Beat Browser 3D [Goldenson, 2007] explores a collection of albums and the songs within those albums. It grew from the Beat Browser project which arranges album covers in a grid in a visual display. In Beat Browser 3D, songs from those albums are placed in a corresponding auditory space. If a particular album is selected, the album "opens" and maps all the songs from the album in a circular space. A mouse is used to select the song or album to listen to. Only one sound source plays at a time, though is synthesised at a specific location in the virtual space.

SoundTorch [Heise et al., 2009] uses the aura model to play multiple concurrent audio files placed on a two-dimensional map. It is designed particularly for audio engineers and designers who need to access sound effects files. It uses both audio and visual information to allow a user to browse through a collection of audio arranged according to similar content. The user directs a cursor over a portion of a two-dimensional map and the audio items within the selected space are sonified over loudspeakers around the user. Finer control of playback is allowed by manipulating the scale of the map and the size of the listening area.

A very literal interpretation of sonifying two-dimensional maps is implemented in City Maps [Heuten et al., 2007]. Maps of physical cities are enhanced with natural sound objects representing physical locations such as bird song in a park. The idea of an auditory torch or aura is again explored, but used in a novel way by allowing multiple and complimentary views of the same space. A user can view the entire auditory space from above or in more detail by walking down the virtual streets.

As graphics cards and three-dimensional game environments became easier to manipulate, two-dimensional maps of songs evolved into immersive three-dimensional environments. Perhaps the most well-known three-dimensional music environment is the nepTune [Knees et al., 2006]. The nepTune adds a visual and audio interface to [Pampalk, 2001] allowing a user to explore a collection of music arranged by similarity. The intended effect is an immersive video game-like environment as clusters of music are placed onto three-dimensional graphic islands rising out of a sea. The soniXplorer [Lübbers and Jarke, 2009] furthers the three-dimensional immersive environment for music exploration as it allows the user to change the environment automatically generated by a self-organising map. Both the nepTune and soniXplorer have multiple audio files concurrently playing from different locations around the user using spatial audio.

The above interfaces are intended for a single user, though as observed in [Cunningham et al., 2003] music browsing and exploration is often a social venture. AudioSquare [Frank et al., 2008] is a multiple-user interactive virtual environment to explore music. Two areas of interest within the virtual world are the musicSOM and Manual Showrooms which display multiple music files arranged according to content similarity or file structure. Multiple users can walk around virtual rooms with virtual loudspeakers playing streams of music.

The AudioSquare and musicSOM interfaces are also extended to an immersive CAVE (cave automatic virtual environment), musicCAVE. The musicCAVE environment displays a three-dimensional interactive model of a map of music with red spheres representing songs. A user can listen to one or more songs by moving towards the spheres in the virtual space. A second interface, which we will refer to as the Social Browser [Adamczyk, 2004], visualises and sonifies a social music network using a variety of visualisation and sonification methods which include a CAVE environment. The same information can also be displayed via a two or three-dimensional visualisation on a computer monitor. Representative music clips are played when a user approaches an artist node in the three-dimensional network and multiple nodes may be audible at the same time. When using the two-dimensional interface a representative music clip is played when a user clicks on an artist node without any spatial audio presentation.

5.1.3 Egocentric perspectives

The *egocentric perspective* or *theater metaphor* described in [Cohen, 1991] consists of a static listener with moving sound sources. The following interfaces use a single listener viewpoint and use different mechanisms to convey the audio content of a collection, though they do not all use spatial audio to do so.

Sonic Radar and Sonic SOM are two complimentary tools for auralizing a two-dimensional arrangement of music [Lübbers, 2005]. Sonic Radar is an audio spatialisation algorithm that introduces a lens element to simulate focus and produces a stereo signal. The Sonic SOM is a visualisation tool of a self-organised map of a collection of music that can be auralized using the Sonic Radar. Sonic SOM generates a map of clusters of songs. For each cluster of songs, a single representative song called a *prototype* is chosen by determining which song is closest to the center of the cluster. The *prototypes* are evenly distributed in a circle around the user in a virtual space. By selecting a song, the user can either launch that song in a media player or explore subclusters within the cluster represented by that *prototype*.

Many auditory displays designed for MIR tasks explicitly aid music browsing and discovery, but do not necessarily focus on efficient presentation of audio content query results. Ali and Aarabi [2008] consider how to return an audio stream representative of query results including rank. We will refer to this work as the Cyclic Interface.

Mused [Coleman, 2007] is another interface that does not use spatial audio, yet has implemented design features that would enhance many MIR tasks. The interface aids composers who use samples in their work by automatically arranging clips of audio into a two-dimensional space. The composer can then mouseover the icons representing each audio clip and hear that audio clip. This allows for passive presentation of audio content as users do not need to intentionally seek out a specific clip but allows for serendipitous discovery.

5.2 Spatialisation Approaches

Synthesising spatial audio can be broadly categorised into two methods: binaural audio for headphones and signals intended for headphone or loudspeaker reproduction. Binaural audio reproduction uses head-related transfer functions, HRTFs, which produce a two-channel signal intended only for headphone playback. Of the 20 interfaces surveyed here, 16 use spatial auditory display. Figure 5.3 shows what spatialisation approaches are used. Six of the interfaces present audio over loudspeakers and seven use headphones. Two of the headphones use stereo panning and five use binaural audio. The remaining three interfaces are unclear in how the spatial audio portion of the interface is

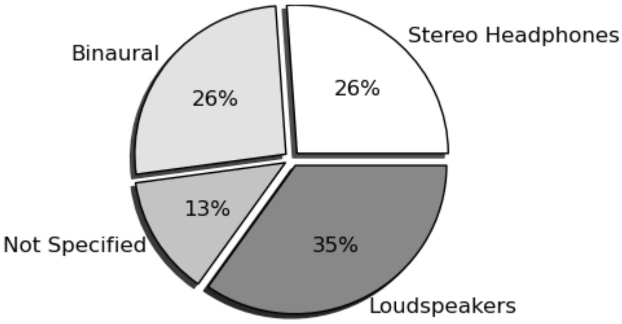


Figure 5.3: Spatialisation approach used for the 17 interfaces with spatial auditory display.

implemented. The City Maps [Heuten et al., 2007] mentions the use of three-dimensional sound and using a three-dimensional sound library with headphones, but never identifies the library nor confirms whether binaural audio is used. AudioSquare [Frank et al., 2008] also never describes how the spatial audio is implemented. It is also unclear for soniXplorer [Lübbers and Jarke, 2009]. The soniXplorer builds on the *focus of perception* described in Sonic Radar [Lübbers, 2005] which is intended for stereo headphone or loudspeaker playback. The playback system for the user evaluation is not described.

5.2.1 Loudspeaker presentation

Loudspeakers are the most common means to place sounds in a virtual sound space, usually using amplitude panning in stereo recordings. Amplitude panning varies the loudness of sound source across multiple loudspeakers allowing the sound to appear to be located between loudspeakers. Synthesis techniques can greatly increase in complexity as the number of loudspeakers increases. While using loudspeakers can avoid some problems commonly encountered with HRTFs, new problems can be introduced. For example, the size of the sweet spot, the area within the loudspeakers where the spatial sound field is reproduced, can vary. Other problems may include the expense and physical placement of a large number of loudspeakers.

Sonic Browser [Fernström and Brazil, 2001, Brazil and Fernström, 2003, Fernström, 2005] places sounds in a spatial sound field using two loudspeakers or a pair of headphones. Audio Retrieval Browser [Brazil et al., 2002], which is based on Sonic Browser, uses the same spatial audio mapping as Sonic Browser. Sonic Radar and Sonic SOM [Lübbers, 2005] also use a stereo signal to pan sources, but they discuss extending to five loudspeakers.

The standard home cinema surround sound system consisting of 5.1 channels is appealing as many video game libraries support it. The nepTune [Knees et al., 2006] and SoundTorch [Heise et al., 2008, 2009] both use 5.1 or 5.0 (no low-frequency effect channel) systems for spatialising audio. MARSYAS3D [Tzanetakis and Cook, 2001] increases the number of loudspeakers to 16 and directly maps an individual audio file to each loudspeaker, though the authors discuss allowing audio files to be panned arbitrarily between speakers in future work.

5.2.2 Stereo headphone presentation

The Playlist Headphones [Hiipakka and Lorho, 2003] intentionally do not use binaural audio because of spatialisation errors, but still use headphones for audio presentation. This is presumably because the interface is intended for mobile devices, in particular mobile phones. The design of the interface is also well-suited for lateralisation, the perception of a sound being located inside the head, and not just localisation, the sound being perceived outside the head. The interface maps menu levels to one of three locations: left, centre, and right.

The Music Scope Headphones [Hamanaka and Lee, 2006] also explicitly use headphones but do not use HRTFs for spatialisation. This again is for portability and accessibility of the interface. The interface is controlled by head movement and hand gestures, so the headphones are also necessary for the mounting of the additional sensors.

5.2.3 Binaural presentation

Binaural is now an attractive choice for spatial auditory display as it has low setup costs. Loudspeaker-based systems can require as few as two loudspeakers but can quickly grow to dozens depending on the spatialisation algorithm. Binaural audio only requires headphones which also makes it an ideal system for mobile applications.

Binaural mobile applications were not a option for the first three interfaces discussed here: Dynamic Soundscape [Kobayashi and Schmandt, 1997], Audio Hallway [Schmandt, 1998], and AudioStreamer [Schmandt and Mullins, 1995]. All three used the Crystal River Engineering Beachtron audio card to render the binaural output. This limited the interface design as only four individual sources could be rendered at the same time. However, the dedicated hardware meant that head-tracking could be performed. Later interfaces (such as [Goldenson, 2007]) are driven by software without dedicated hardware and do not have head-tracking as it is an added computational expense and requires specialised hardware. The Social Browser's three-dimensional presentation does use head-tracking [Adamczyk, 2004].

Binaural audio is prone to errors when non-individualised HRTFs are used. When a listener listens to binaural audio that was generated with HRTFs that are not their own, the HRTF set is non-individualised. The severity of errors introduced by non-individualised HRTFs varies according to the individual listener and HRTF set. The authors of the Audio Hallway [Schmandt, 1998] noted that localisation errors may have been exacerbated by using HRTFs from a dummy head and using source audio with a bandwidth of only 4kHz. In general, broader bandwidths are easier to localise [Blauert, 1997].

5.3 Two-Dimensional Maps and Three-Dimensional Worlds

As MIR techniques improve, greater insight can be made into a collection of audio, in particular, how different audio files relate to each other. MIR researchers often add spatial

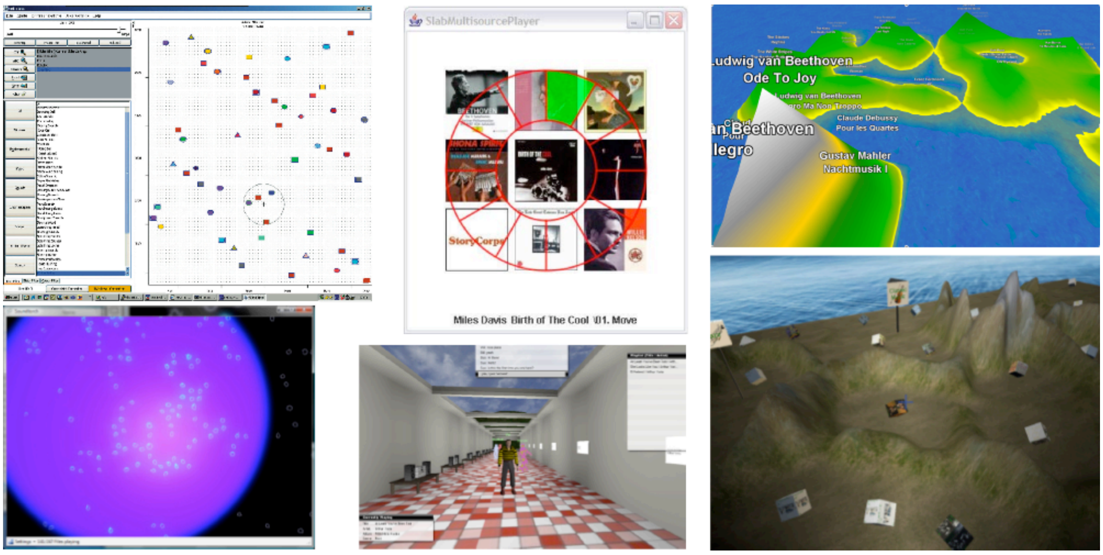


Figure 5.4: Three two-dimensional maps and 3 three-dimensional worlds. The top row left to right: Sonic Browser [Fernström, 2005], Beat Browser 3D [Goldenson, 2007], nepTune [Knees et al., 2006]. The bottom row left to right: SoundTorch [Heise et al., 2009], Audio Square [Frank et al., 2008], soniXplorer [Lübbers and Jarke, 2009]

elements to representations of audio collections when trying to succinctly represent a large amount of data. Most visualisations of these relationships create two-dimensional maps representing music collections.[1] A natural extension to visually examining generated maps of songs is to sonify the data and attempt to listen to how the maps "sound" with the hope of learning more about the represented audio. This usually means utilising spatial audio to create an auditory display or a multi-modal interface with some combination of audio, visual and haptic interaction.

There are 11 interfaces that provide direct sonification of a two-dimensional or three-dimensional mapping of audio files (all 11 are listed in Figure 5.2). While some of them have specific mapping mechanisms, they all can be used to interact with any generic map of audio files. Sonic Browser, MARSYAS3D, the Audio Retrieval Browser [Brazil et al., 2002], and Beat Browser 3D [Goldenson, 2007] either do not use a specific mapping mechanism or can interchangeably use a number of maps.

Beginning with Sonic Browser [Fernström and Brazil, 2001, Brazil and Fernström, 2003, Fernström, 2005], two-dimensional arrangements of audio for visual inspection of trends and information began to be sonified. Users listened to the audio content of multiple audio files spatialised at different locations. The locations reflected the arrangement of icons on a visual display.

The factors driving the arrangement of audio files vary in complexity. By far the most popular mechanism is the self-organising map (SOM) which drives Sonic SOM [Lübbers, 2005], nepTune [Knees et al., 2006], the MusicSOM Showroom within AudioSquare [Frank et al., 2008], SoundTorch [Heise et al., 2009], and soniXplorer [Lübbers and Jarke, 2009]. SOMs are neural network structures that are usually used in music to group

[1] See http://visualizingmusic.com/

similar-sounding audio files together on a two-dimensional map. The map is calculated using feature vectors extracted from the content of the audio files.

City Maps [Heuten et al., 2007] displays the geographical map of a city to blind users, so the mapping of sounds is quite literal as representative auditory icons are used to identify landmarks. Most of the interfaces using a map paradigm have visual displays with the exception of City Maps. A number of the two-dimensional maps and three-dimensional worlds that have corresponding visual displays can be seen in Figure 5.4.

Figure 5.2 shows that the majority of the interfaces sonifying a two or three-dimensional map also use spatial audio in that sonification. The exception is Mused [Coleman, 2007] which plays back a single audio clip at a time. A user only needs to move a mouse over the icon of a music clip for the clip to play, allowing for quick succession of clips.

The first person narrative of exploring a map of music is easily exploited by video game paradigms complete with three-dimensional graphics. The first such interface is nepTune [Knees et al., 2006] followed by AudioSquare [Frank et al., 2008] and soniXplorer [Lübbers and Jarke, 2009].

5.3.1 Auditory landmarks

When interacting with a map of music, users need to retain information about the virtual space and not become disoriented, but music is a time-based medium. This means that audio files may need to be limited in time so that a particular sound or auditory landmark can be easily associated with a location in the virtual space. The nepTune [Knees et al., 2006] does this by limiting the audio playback to the middle 30 seconds of each song so that the music does not change too much if a user returns to a previously-visited virtual position. The Social Browser [Adamczyk, 2004] is not strictly a two-dimensional or three-dimensional map as it is a displayed network of social linkings of similar artists, but it limits clips to 20 seconds.

Similarly, the MusicCAVE implementation of the AudioSquare environment [Frank et al., 2008] has a return to 'home' function so that users can return to a known location if they become disoriented within the virtual space. The soniXplorer [Lübbers and Jarke, 2009] does not use auditory landmarks, but allows users to add visual markers to the three-dimensional landscape. The Beat Browser 3D [Goldenson, 2007] employs a different view on auditory landmarks and instead keeps a global clock, so that all tracks are continually playing (though only a select few are heard at any time). The intent is to allow browsing through albums.

5.4 Navigation Tools

"Overview first, zoom and filter, then details on demand" [Shneiderman, 1997] is an often quoted design principle behind many of the interfaces discussed here. It is difficult to allow for an overview of a collection of audio content, but often multiple audio files are presented concurrently to more efficiently scan a collection of audio. The approaches to control the presentation of concurrent audio streams is first discussed along with how

those controls allow for zooming. Filtering content so that only the most pertinent data is presented is then explored. This section concludes with additional tools that are frequently employed in auditory displays and how customisation and adaptation allows interfaces to conform to the user.

5.4.1 Controlling multiple streams

A primary concern expressed by users of music interfaces during evaluations is that listening to more than one sound source can be confusing and that they can be quickly overwhelmed with too much information. There are limits on the number of sound sources that a user should be presented with. Lorho et al. [2001] found that users could accurately describe sources best when they were spatially separate with different onsets and limited to three sources. The MARSYAS3D system [Tzanetakis and Cook, 2001] could play up to 16 sound sources each mapped to a single loudspeaker, but it was found that a maximum of eight sources was best.

The Sonic Browser coined the term *aura* for a means to restrict the number of sound sources "a function that indicates the users range of interest in a domain" [Fernström and McNamara, 1998]. It is commonly visually represented as a circle as seen in a number of the visual displays in Figure 5.4. The function is to limit the number of audio files that play back concurrently and to highlight a region of interest. This idea was again repeated in the SoundTorch [Heise et al., 2009], and also as an "auditory torch" in City Maps [Heuten et al., 2007].

Auras or functionally similar tools are frequently requested if an interface with multiple concurrent streams of audio does not contain an aura [Fernström and Brazil, 2001, Adamczyk, 2004]. Users testing nepTune [Knees et al., 2006] asked for "larger landscapes to allow focused listening to certain tracks in crowded regions." None of the three video game-like interfaces have an adjustable aura, but soniXplorer does try to limit the number of audible sound sources by not allowing multiple songs to play from the same location. If multiple songs are located near each other, only the closest one will be heard.

AudioStreamer [Schmandt and Mullins, 1995] takes a different approach to zoom and filter. It enhances the user's focus by increasing the gain of the source that the user has turned their head towards by 10 dB. To mimic decreasing interest, the gain of the focussed sound decreases over time, but the user can register that they are retaining interest by turning towards the source again. Once interest has been registered three times, the other sources are silenced.

This same idea of focusing is also used in the rooms in the Audio Hallway [Schmandt, 1998]. Each room contains a cluster of semantically-related audio clips which are arranged around the user's head. The audio the user is focussing on is louder than the surrounding songs to allow them to pay closer attention to that source.

The Sonic Radar alludes to the focusing of a lens on a particular audio file while other files play in the periphery [Lübbers, 2005]. A song selected from the focus of the interface could launch that song in a media player or, if it was a *prototype*, allow the user to explore the sub-cluster, giving “details on demand”. Music Scope Headphones [Hamanaka and Lee, 2006] also uses an increase in volume to emphasise the audio information located in front of the user.

Restricting where information can occur may also decrease confusion. Humans are best at localising sounds in front and to the sides of the head, but are poor at localising sounds above and behind the head [Begault, 2000]. The Playlist Headphones [Hiipakka and Lorho, 2003] compensate for this by limiting the placement of virtual sounds to three positions while Music Scope Headphones [Hamanaka and Lee, 2006] allow the user to restrict where around the head songs can play. Users can also choose for themselves the source locations around their head if they wish to only listen to a subset of the collections. They can also easily return to the full set of presented songs.

5.4.2 Abstraction, segmentation, and compression

Giving the user an *overview ... then details on demand* is not a simple task with time-based media such as audio. This is usually addressed in video with representative stills and can be extended to audio with short, representative clips of audio called thumbnails [Aucouturier and Sandler, 2002], but several interfaces took an even more innovative approach. The AudioHallway uses a technique described as *braided audio* to give the user an overview of the content inside a virtual room containing a cluster of semantically-related audio by rotating playback through various 3 second clips from the cluster [Schmandt, 1998]. These audio braids are then placed in the virtual space as doors along a hallway to allow the listener to have a broad representation of the audio content of the virtual room.

SpeechSkimmer [Arons, 1993, 1997] has four skimming levels or levels of compression of speech content referred to as a “fish ear” (as opposed to a fisheye lens). This is to enable a user to quickly scan the content and then gradually slow down the scanning when an area of interest is identified. The four skimming levels are not directly applicable to musical content, but the principle is for the lowest level to be unprocessed audio. The next two skimming levels shorten or remove pauses so that the audio is unnaturally fast, but without any changes in pitch or in comprehension. The top two skimming levels attempt to automatically segment the audio into semantically-related sections so that quick scanning of each segment can occur.

Zooming into finer detail is also implemented in the Music Scope Headphones [Hamanaka and Lee, 2006]. Multiple songs are initially presented to the user, but a single song can be selected. That song is then separated into the individual instrument parts and arranged in space. The user can then interact with them as if listening to multiple songs.

While the Audio Hallway derives a hierarchy by clustering similar audio files and then providing summaries of those clusters, Playlist Headphones [Hiipakka and Lorho, 2003] use the traditional artist-album-track tree for navigation of audio files. However, MIR research can derive far more complex relationships from audio content and metadata than this traditional structure. These relationships are often arranged in a two-dimensional map to allow for visualisation of the data. Many of the interfaces sonify two-dimensional maps of audio files that are arranged in a flat hierarchy ([Fernström and McNamara, 1998, Tzanetakis and Cook, 2001, Lübbers, 2005, Hamanaka and Lee, 2006, Knees et al., 2006, Heise et al., 2009, Goldenson, 2007]), but this does not necessarily provide an *overview . . . then details on demand.* Perhaps the data can benefit from being slightly obfuscated behind more structure. Sonic Radar [Lübbers, 2005] takes a different approach by creating a hierarchy from the clustering on the map by choosing a *prototype* song to represent a cluster. This gives the two-dimensional map more order and allows the listener to have a greater overview of the map without being overwhelmed by concurrently playing audio content.

Rather than interacting with individual songs, some interface interact with artists. The Social Browser [Adamczyk, 2004] is an example of such an interface and uses representative tracks of each artist chosen by the authors. However, this is not a scalable solution. Even when songs do not need to further represent a concept like an artist, it is difficult to automatically identify the best portion of the song to present. Many authors choose to select an arbitrary excerpt [Adamczyk, 2004, Knees et al., 2006]. As suggested by the authors in [Tzanetakis and Cook, 2001], automatic thumbnailing is an option, though none of the interfaces described here use it.

The Cyclic Interface [Ali and Aarabi, 2008] concatenates multiple audio clips into a single stream and weights each individual audio file to reflect its ranking as a result to a query (no spatialisation is used). The effect is song clips fading out as another clip fades in so that multiple songs can be quickly reviewed. A visualisation of the audio provides metadata and helps the user to differentiate between songs. The audio signal is analyzed in order to automatically present the sections of the song that may be of the most interest. The sections with the highest level of information in the frequency domain, high spectral entropy, are considered the most interesting. Only the first 20 to 30 seconds are considered.

5.4.3 Auditory icons and text-to-speech

While the interfaces discussed are reliant on auditory display as a means to convey information, most of them have complementary visual displays. The SpeechSkimmer [Arons, 1993, 1997], Audio Streamer [Schmandt and Mullins, 1995], AudioHallway [Schmandt, 1998], Playlist Headphones [Hiipakka and Lorho, 2003], Music Scope Headphones [Hamanaka and Lee, 2006] does not a have corresponding visual display and relies on only audio information. The Playlist Trackball [Pauws et al., 2000] compares the same auditory display and haptic interface with and without a visual display.

An auditory display which sonifies audio data such as a digital music uses direct sonification, but most auditory displays for non-music applications utilise abstract sounds to represent actions or states of the interface. Auditory icons and earcons are two formal non-speech approaches [Gaver, 1986, Blattner et al., 1989]. Auditory icons are "caricatures of naturally occurring sounds" that are intended to bear some direct relation to what they represent [Gaver, 1986]. Auditory icons are a subset of earcons and are referred to as representational earcons. Abstract earcons are sounds not associated with a physical action; "[a]rt is to icons as music is to earcons"[Blattner et al., 1989].

Auditory icons are important for conveying functionality with an interface, but are used in only one of the auditory displays discussed here: the Playlist Trackball [Pauws et al., 2000]. Metallic rolling and hitting sounds are used to give feedback regarding the physical trackball interface. Text-to-speech is slightly more common as it is used in the Playlist Trackball and also the Playlist Headphones [Hiipakka and Lorho, 2003] to read items such as artist names and menu titles.

5.4.4 Customisation

Sonic Browser, MARSYAS3D [Tzanetakis and Cook, 2001], and the Audio Retrieval Browser all allow for easy re-mapping of a collection of audio files onto a two-dimensional or three-dimensional space according to varying parameters, examples of which are file size, timbre, and tempo. This allows for a custom perspective of the collection depending on the aspect the user is interested in. In a similar manner, with the Playlist Headphones [Hamanaka and Lee, 2006] a user can switch between a number of presets that determine where songs are located in space.

Multiple perspectives of the same data are also allowed in City Maps but without remapping the audio files. In City Maps audio files are mapped to static locations on a geographical map. The user can explore the map either by virtually walking on the map with audio files distributed on a horizontal plane from a first-person perspective or by 'listening as a bird' and exploring the map on a vertical plane in front of the user. The user then uses an aura in a similar manner as SoundTorch to sonify the map.

Multiple views are not offered in soniXplorer [Lübbers and Jarke, 2009], but a user can alter the presented map of songs. Users can move songs that they feel are misplaced. They can also change the terrain by adding or removing hills so that songs are more or less partitioned. They can also add visual landmarks to the terrain to indicate a location of interest. The underlying system then learns these changes, adapts the landscape and new songs to attempt to suit the user's tastes.

5.5 Usability

Of the 20 interfaces examined here, 12 have published a formal evaluation. A formal evaluation is defined as one with enough detail to report the number of participants. The ubiquitous informal evaluation is found with most of the interfaces not described in Table 5.1. However, it is difficult to draw substantial conclusions from anonymous, uncounted interface testers.

Table 5.1 gives an overview of the 12 interfaces with usability evaluations. It highlights the number of participants in each usability test, the evaluation techniques used, whether an experimental control was used, and if any statistical analysis was performed on the results. Five of the studies provided basic quantitative summary information, though only the Playlist Trackball [Pauws et al., 2000] tested for statistical significance.

The depth of evaluation is fairly dependent on where the interface was published.

Interface	Users	Evaluation Technique(s)	Control Task	Stat. Analysis
SpeechSkimmer Arons [1997]	12	think aloud, questionnaire, interview, videotape	X	
Dynamic SoundScape Kobayashi and Schmandt [1997]	4	task (measured accuracy)		
Playlist Trackball Pauws et al. [2000]	24	task (timed and measured accuracy), questionnaire	X	X
Sonic Browser Fernström and Brazil [2001]	6	think aloud, task	X	
Playlist Headphones Hiipakka and Lorho [2003]	10	task (measured accuracy), questionnaire		
Social Browser Adamczyk [2004]	6	think aloud, interview	X	
Music Scope Headphones Hamanaka and Lee [2006]	3	task (timed)	X	X
nepTune Knees et al. [2006]	8	interview		
City Maps Heuten et al. [2007]	17	not described		
Cyclic Interface Ali and Aarabi [2008]	20	task (timed), questionnaire	X	X
SoundTorch Heise et al. [2008]	15	task (measured accuracy), questionnaire	X	X
soniXplorer Lübbers and Jarke [2009]	9	task (measured accuracy), questionnaire	X	X

Table 5.1: Comparison of user evaluations of interfaces. If an interface is not listed, a formal evaluation was not found. Only evaluations with enough detail to state the number of participants have been included. A study is deemed to have a statistical analysis if any numerical summary information is provided, though only Playlist Trackball [Pauws et al., 2000] tested for statistical significance.

Understandably, the interfaces primarily published in the HCI literature are more thoroughly evaluated, with SpeechSkimmer [Arons, 1997] and the Playlist Trackball [Pauws et al., 2000] having the two most extensive evaluations. It is slightly alarming that there appears to be a trend to not thoroughly perform evaluations past 'informal evaluations' or 'heuristic design.' While informal techniques are necessary for iterative design cycles, much is to be gained from thorough testing with potential end-users [Jeffries and Desurvire, 1992].

SpeechSkimmer [Arons, 1997] cited using informal heuristic evaluation throughout the design process, but it was also evaluated by 12 people. The participant, an observer, and an interviewer were present for each trial, and it was video-taped throughout. The researchers were explicitly concerned with how easy the interface was to use without prior instruction and how it compares to the 'traditional' approach. Each trial was about 60 minutes long and was broken into five parts:

1. A background interview was undertaken to determine the participant's prior experience with recorded speech and audio.

2. The participant was given the physical interface and asked to describe how it might be used without guidance from interviewer.

3. The participant freely used and explored the device and was encouraged to think aloud while doing so.

4. A comparison exercise was conducted by having the participant access the same content segmented in two different ways but using the same interface. The participant judged both types of segmentation using a 7 point scale and then answered 3 questions that could be easily answered by listening to the audio content. The participant then described the search strategy used to answer the questions. The audio content and order presented were varied so that all combinations were covered.

5. A follow up interview gathered the participant's opinions and thoughts concerning the evaluation.

The authors of the Playlist Trackball [Pauws et al., 2000] were interested in the effect of a visual display and instant usability or how much a user understands without explicit instruction. The authors were very interested in quantitative evaluation of the interface and worked to produce a statistically valid study. There were 24 participants who participated in two sessions on two different days. They were randomly assigned to one of four conditions: four were assigned the control condition using the full multimodal system with visual display for all four tasks; four were assigned the control condition with no visual display for all four tasks; eight were assigned the experimental condition with visual display for the first two tasks and no visual display for the third and fourth task; and eight were assigned the experimental condition with no visual display for the first two tasks and visual display for the remaining two. Each session consisted of:

1. The participant was given 15 minutes to become familiar with only the physical interface and some general instruction on building playlists.

2. 3 minutes to explore full interface with or without visual display accordingly.

3. The participant performed two playlist building tasks. At the beginning of each task, the participant was given written instructions which they had to rewrite in their own words to demonstrate they understood.

4. The participant completed a questionnaire after each task. The questionnaire contained the same questions but in a randomly different order each time. The questions covered how the interface is used, so answers were either correct or incorrect.

Pauws et al. [2000] examined the number of actions required to complete each playlist task and scores from the questionnaires. They then analyzed the results using ANOVA (analysis of variance) and MANOVA (multivariate analysis of variance) tests to determine whether visual displays and practice had any measurable effect. Hypotheses were then accepted or rejected accordingly.

These two experiments are examples of good experimental design. They both state hypotheses or research questions (easy to learn without assistance, effect of visual display, etc.). Experimental controls and randomisation minimise bias while also using statistical analysis where appropriate to determine the significance of the results. Both also use a large number of participants in comparison with other studies as is seen in Table 5.1; the number of participants in usability tests ranges from 3 to 24 with a median of 10.

The time a participant is allowed to become familiar with an interface is important, especially if the interface is intended to have instant usability as is stated in [Arons, 1997, Pauws et al., 2000, Hiipakka and Lorho, 2003]. Even if instant usability is not a pertinent research question for a given interface, it is important to ensure each participant has the same exposure to the interface during evaluation. On average, participants are given about 5 minutes.

While general aesthetics and other qualitative aspects of an interface are important, task-based evaluation can concisely measure the effectiveness of an interface. As most of the interfaces discussed here are intended to replace or enhance established music search or browsing tools, the proposed interface can be used to perform the same task as an established interface. By using a control task, SpeechSkimmer [Arons, 1997], Sonic Browser [Fernström and Brazil, 2001], Music Scope [Hamanaka and Lee, 2006], SoundTorch [Heise et al., 2008], and the Cyclic Interface [Ali and Aarabi, 2008] have shown that they are more efficient tools for music search and retrieval than traditional interfaces without auditory display. Pauws et al [Pauws et al., 2000] have demonstrated that a haptic interface with auditory and visual displays is equally as effective without any visual display; users are only slower but not more error-prone.

5.6 Summary

Listening to audio content has demonstrable effects on how people interact with a collection of music [Barrington et al., 2009]. The Cyclic Interface [Ali and Aarabi, 2008] shows that users do perform search tasks more efficiently when audio in presented in a non-serial manner. It has repeatedly been shown that when audio is integrated into a browsing or search system, users are able to find content much more efficiently [Arons, 1997, Fernström and Brazil, 2001, Hamanaka and Lee, 2006, Heise et al., 2008].

Early spatial auditory display for navigation of audio files initially looked at interfaces without any visual display and used binaural audio to present information. Hardware constraints limited the auditory display as few virtual sources could be simultaneously rendered and the interfaces were confined to research laboratories. However, these restraints also allowed for innovative developments by introducing gestural control. Sophisticated interactive computer graphics were not available, so more complex audio representations were explored. Much of the research in the 1990s looked at methods to segment and abstract data so that a broad understanding of the content could be made and specific details easily retrieved.

As MIR visualisation techniques progressed, sonification of those visualisations followed. The seminal Sonic Browser [Fernström and Brazil, 2001, Brazil and Fernström, 2003, Fernström, 2005] inspired the design of many interfaces for two-dimensional arrangements of audio collections. As hardware became more portable, powerful, and flexible, interfaces began to be developed for mobile devices and also for immersive applications. Binaural audio and headphone reproduction is now being utilised in spatial auditory displays for devices with little or no visual information. Additionally, large numbers of loudspeakers are also being used to create auditory displays to complement computer graphics in a video game-inspired or immersive CAVE environment.

Auditory display has been used for a variety of music tasks, namely to build a playlist [Hiipakka and Lorho, 2003, Pauws et al., 2000], music browsing or search [Hamanaka and Lee, 2006], and to efficiently present search results [Ali and Aarabi, 2008]. These tasks are not independent and users often need to easily move between targeted searches and exploratory browsing [Cunningham et al., 2003]. However, common auditory display design features continually reemerge irrespective of the specific task for which the interface was designed. These include: multiple streams of audio playing concurrently; auras to limit the amount of presented information and highlight a particular selection; two-dimensional or three-dimensional mappings of an audio collection to a virtual space; user customisation of virtual spaces to tailor the interface to their needs; and a non-flat structure or hierarchy along with thumbnailing of the audio content to allow for an accessible overview of a large collection of time-based media.

Spatial auditory display relies on affecting cues such as ITD and ILD to determine the angle and elevation of a sound source. Still angle and elevation are only a part of sound source localisation; distance completes the soundscape. Reverberation plays a key role in distance perception, in particular with regard to the effect of the ratio of direct to reverberant sound [Begault, 2000]. None of the interfaces discussed here explicitly mention reverberation in the auditory design, though it may have been a feature used in the sound libraries used to render the auditory display. Reverberation was not an option in early interfaces [Arons, 1997, Schmandt and Mullins, 1995, Schmandt, 1998, Kobayashi and Schmandt, 1997] due to hardware constraints. Reverberation is often avoided in auditory displays as it decreases localisation accuracy. However, if appropriate for the task, reverberation can create a more realistic, immersive environment. Additionally, while reverberation will decrease the user's ability to correctly determine the directions of a sound source when compared with a virtual environment without any reverberation, it has been shown that listeners can "learn" the reverberation over time and become better at localisation tasks [Shinn-Cunningham, 2000]. Precise localisation of a sound source is most likely not necessary in the majority of auditory display designed for music search or browsing tasks, but an immersive and pleasant audio experience is important.

Figure 5.1 shows that auditory displays for music or speech tasks were initially published within HCI communities where usability tests are expected and evaluation techniques are well-understood. As interfaces began to be published in complementary communities such as signal processing and MIR research publications, thorough usability tests have diminished, but not vanished. Of the 20 interfaces, 12 conducted a usability study, though as seen in Table 5.1 most of those studies could further embrace HCI evaluation techniques and improve the use of statistics to validate whether results are significant.

The next challenge to researchers utilising spatial audio for music tasks is to let previous interfaces inform future design. Early interfaces had to use large dedicated digital signal processors to handle rendering even simple scenes, but even with limitations in computing, the interfaces had elegant designs. Now that spatial audio and computing power have advanced, perhaps an even better marriage between technology and design can emerge.

After observing people searching through music collections in music stores and libraries, Cunningham et al. [2003] conclude that music discovery is a very social event. Only one interface, Audiosquare [Frank et al., 2008] allows for direct social interaction. Perhaps future interfaces should further explore how direct sonification and direct social interaction can improve music discovery.

Cunningham et al. [2003] emphasised that searching, browsing, and activities such as building a playlist are not mutually exclusive events, but are complementary processes used to access and interact with a collection of music. The map interfaces tend to be good for browsing and exploration while the cyclic interface and playlist-building interfaces are well-suited for navigating results of a retrieval query or search task. A user needs to perform both types of tasks and often interchangeably. Future interfaces need to reflect this and as has been shown here, spatial auditory display is a highly effective approach to improve interaction. Spatial auditory display will be a key component in such a system.

Chapter 6

The amblr: A Novel Music Browser

As music collections grow in size, better tools are needed to manage the content. Text-based tools for finding music are the most common as most music content systems such as iTunes organise music into artist/album/song hierarchies using textual metadata. Interfaces to online services with millions of tracks such as last.fm, Spotify, and the iTunes Music Store also rely on text-based searching, but text is an inadequate means to describe music as concepts like timbre are difficult to express verbally. Additionally the text attributed to a song may not be correct or could just be misspelled.

Text is most commonly used in artist, album, and song title, but it also is used as tags which are words or phrases to describe a song or artist. When text is used as tags, the meaning of a word or phrase is often ambiguous and has a high variance in interpretation, but text can also be limiting. When text such as an artist name is used to recommend music, only 'well-known' music will be further recommended [Celma and Cano, 2008]. Furthermore, familiarity with song's metadata affects a user's reaction to a recommendation; if they are presented with the textual description of a song, such as artist and song title, they will respond differently than if they are required to listen to the audio content of the recommendation [Barrington et al., 2009].

In most current interfaces text is the primary means of navigating a collection or discovering new music. However, if the only means to navigate a collection of music is by textual metadata and that metadata becomes corrupted or simply holds no meaning to a particular user, navigation of the collection is broken, and all or some of a collection becomes inaccessible. Subsequently, text can be viewed as a series of barriers to accessing the audio content of a collection of music, making serendipitous browsing difficult.

As demonstrated by [Adamczyk, 2004, Brazil et al., 2002, Coleman, 2007, Fernström and Brazil, 2001, Frank et al., 2008, Goldenson, 2007, Heise et al., 2009, Knees et al., 2006, Lübbers, 2005, Lübbers and Jarke, 2009, Tzanetakis and Cook, 2001], a common approach to exploring a collection is to create a virtual space populated by music. Of the above 11 interfaces, all but Mused [Coleman, 2007] use spatial audio. With Mused a user moves the mouse over the icon of a music clip for the clip to play, allowing for quick succession of clips to be played, but like Beat Browser3D [Goldenson, 2007], only one audio file at a time is played. The other 9 interfaces allow multiple audio files to play concurrently, but separated spatially.

Considering the previous work surveyed in Chapter 5, how can audio be used to facilitate music discovery and what design features should continue to be used and what new ones should be introduced? This chapter presents a new spatial audio interface for exploring a collection of music: the amblr. The amblr is a step forward in music browsing, removing the over-reliance on textual metadata and visual displays. It builds on established approaches from previous work and pulls together disparate auditory display tools while incorporating additional novel design features. Using spatial audio to explore spatial arrangements of audio collections is a continually revisited design, but user tests show that the previous interfaces are not entirely intuitive or a satisfactory solution for exploring a collection of music. There are some approaches that recur in multiple interfaces, while other interfaces use unique approaches that have not been repeated.

The amblr is inspired from previous music browsers that interact with two-dimensional maps of songs. The amblr builds on: auras first described in [Fernström and Brazil, 2001]; perception of focus as implemented in [Lübbers and Jarke, 2009]; auditory landmarks as discussed in [Adamczyk, 2004]; and cyclic presentation of audio content as explored in [Ali and Aarabi, 2008]. Additionally, the amblr utilises audio confirmation of a selection, haptic interaction, and an auditory display that does not rely on a visual display. The amblr does not accommodate all of the suggestions made in Chapter 5 in order to simplify the interface. Social interaction, non-flat hierarchies, and user-customisation of the virtual space are not supported and only a basic spatial audio environment without any reverberation is created. Instead the amblr only investigates how to build directly on the two-dimensional map paradigm that has been so often implemented.

This chapter follows the design and evaluation of the amblr beginning with the first design iteration and user evaluation in Section 6.1. Results from the user evaluation inform the second design iteration described in Section 6.2, and Section 6.3 describes a public art installation based on the amblr.

6.1 First Design Iteration

The amblr is developed in two design iterations. The first design iteration establishes an initial proof-of-concept which is evaluated by 12 users. It is necessary to define initial parameters that the interface needs to satisfy. The initial design criteria are as follows:

- No visual display
- No hardware beyond headphones and mobile device
- Gestural control and haptic feedback using buttons and sensors available on mobile device
- No head tracking, let user move in virtual space to resolve front-back confusion
- Two-dimensional virtual space populated by music that is not restricted to any particular space
- Binaural audio rendered with virtual Ambisonics

6.1.1 Interface design

The amblr's interface assists a user in selecting a song from a large collection of music without any visual feedback. It is designed as a mobile application, so it assumes little or no visual information can be displayed and that headphones are the most convenient means to present audio.

The amblr has no knowledge as to how the songs of a music collection are arranged in a two-dimensional space. Only the coordinates for each song in the collection and the location of the user are known. This allows the interface to be used in conjunction with any playlist generator or similarity map that can generate a unique two-dimensional coordinate for each song such as those discussed in Section 5.3.

Few of the interfaces discussed in Chapter 5 discuss how spatial audio is rendered. The earlier interfaces such as [Kobayashi and Schmandt, 1997, Schmandt and Mullins, 1995, Schmandt, 1998] were limited by processing constraints as dedicated hardware was necessary to compute binaural output. The design was then restricted as the number of concurrent sound sources was limited. More recent interfaces such as [Knees et al., 2006, Lübbers and Jarke, 2009] make use of video game development platforms in order to integrate 3D graphics. However, game platforms intentionally obfuscate how audio scenes are rendered to simplify development. The programmer can only control where a sound is placed in a virtual environment and cannot control how the spatial audio is rendered.

Dedicated audio processing hardware is neither necessary nor desirable for a mobile application, and game platforms do not easily allow for customisation of audio rendering. Therefore the amblr is built on virtual Ambisonics as described in Chapter 3. Virtual Ambisonics is advantageous over convolution with HRTFs because: an increase in the number of concurrent playing sound sources does not significantly increase the processing load; binaural audio uses headphone playback; and head tracking and other rotational movements are simple in the B-format domain. Head-tracking is not implemented here as it requires additional hardware beyond the headphones and computer processing the audio environment.

Non-individualised HRTFs commonly create front-back confusion for many listeners with head tracking being an effective method to overcome the errors [Begault et al., 2001]. However, the equipment is limited to headphones and a processing unit (a computer) with the aspiration to use this interface in mobile applications with minimum equipment. The intent is that users will be able to resolve some front-back confusion and other localisation problems by moving in the environment and manually changing the directional cues they receive.

The amblr builds on previous work which uses two-dimensional maps to explore a music collection. All of the previous interfaces using two-dimensional maps have visual displays showing some amount of information about the music collection. Audio is kept as the primary mode, so there is not a visual display. Haptic interaction is used to manipulate the interface by using sensors that are becoming more commonly integrated into mobile devices such as phones: accelerometers, buttons, and vibrating motors or batteries.

6.1.1.1 Navigating the map

The most basic interaction with a two-dimensional map of music is allowing the user to move virtually around the map. The user occupies a location on the map along with an orientation indicating the forward direction of the user. They can move forwards, backwards, left, right, or any combination of those directions to change locations. However, the user cannot change their orientation, so they cannot rotate the sound field. This is to simplify the gestures needed to move around the map as will be discussed in Section 6.1.2.

By moving around the map the user can approach or move away from songs as if they are walking around the map. While they may hear multiple songs at the same time, they can control the volume of each song through their proximity to each song in the virtual space. All of the songs exist on the same plane as the user's head; songs are not above nor below the user.

6.1.1.2 Zoom

During user evaluations of similar audio environments, users have been concerned that listening to more than one sound source can be confusing and that they can be quickly overwhelmed with too much information. See Section 5.4.1 for a further discussion. Similarly, during the development of the amblr, merely describing an interface with concurrently playing sounds to potential users elicited skeptical comments about the approach. Users are nervous about information overload and need a mechanism to throttle the presentation of information.

The *aura* first described by Fernström and McNamara [1998] has become a popular approach to limiting the number of audio files that play back concurrently. This idea has been repeated as an 'auditory torch' in City Maps [Heuten et al., 2007] and in SoundTorch [Heise et al., 2009]. The amblr refers to the functionality of an aura as zooming as it is similar to zooming in and out of an online geographical map interface. Zooming is illustrated in Figure 6.1.

The user can hear only the songs within the surrounding listening area. When zooming in, the listening area shrinks so only the closest songs can be heard; when zoomed out, the listening area grows allowing more songs to be heard. A larger listening area may be beneficial when trying to gain an overview of the map, but may be too much information if searching for a particular song. The user zooms in to listen to only the songs closest to the listening position.

6.1.1.3 Auditory landmarks

When interacting with a map of music, users need to retain information about the virtual space and not become disoriented. This is particularly important when there are no visual cues to guide a user through a space. However, music is a time-based medium so the audio playing at a given location may be different at another point in time. This means that audio files may need to be limited in time so that a particular sound or auditory landmark can be easily associated with a location in the virtual space.

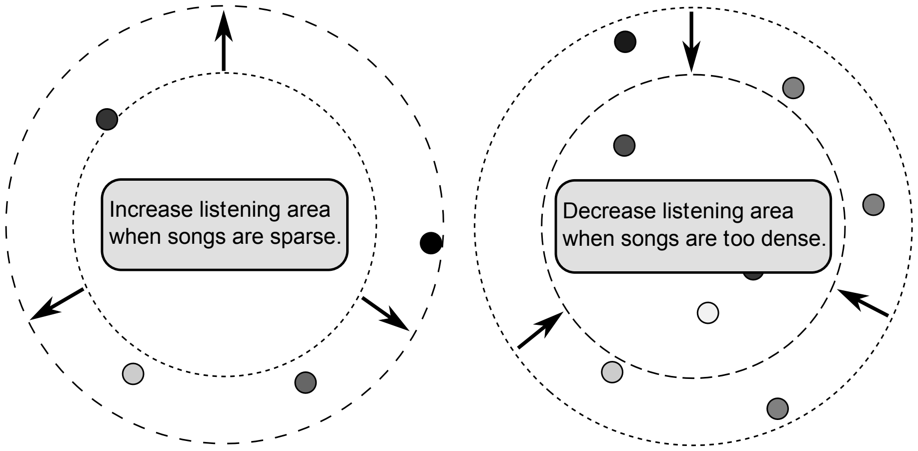

Increase listening area when songs are sparse.
Decrease listening area when songs are too dense.

Figure 6.1: Illustration of the how the zoom function can be used to navigate through dense or sparse data.

The nepTune [Knees et al., 2006] does this by limiting the audio playback to the middle 30 seconds of each song so that the music does not change too much if a user returns to a previously-visited virtual position. The Social Browser [Adamczyk, 2004] is not strictly a 2D or 3D map as is a displayed network of social linkings of similar artists, but it limits audio clips to 20 seconds. While automatic thumbnailing attempts to identify a short representative clip of a longer piece of music, it is not a solved problem [Aucouturier and Sandler, 2002]. The data set discussed in Section 6.1.4 consists of audio clips approximately 30 to 90 seconds long. This is the full length of the provided audio file which is the song selection made freely available on music purchasing sites. The amblr plays the full audio clip.

The MusicCAVE implementation of the Audio Square environment [Frank et al., 2008] has a return to 'home' function so that users can return to a known location if they become disoriented within the virtual space. The amblr contains this feature as well. When the amblr is requested, the user is returned to the centre of the map. The zoom level is not changed since there is no way to notify the user of a change in the zoom level; the zoom level can only be explicitly altered by the user. If a user becomes disoriented while exploring the map of music, they can return to the 'home' position where the virtual space should be familiar.

6.1.1.4 Selection

The amblr is intended for music browsing. If a song of interest is found, then a mechanism to select that song is needed. The selection of a song could be used for a variety of tasks such as becoming the seed to a search, being added to a playlist, or perhaps purchasing that song from an online store. All of those tasks require the browsing of a collection of songs and then the selection of a single song.

As there is no visual confirmation of the selected song within the amblr, an aural confirmation is presented. The user selects a song by approaching the song in the virtual space. To confirm their choice, after indicating that they would like to select it, they then listen only to the selected song. While the interface spatialises each song from a unique

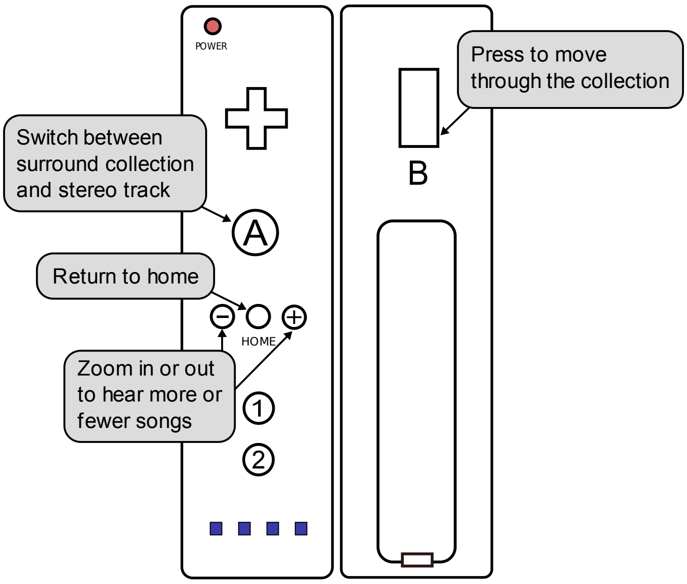

POWER
Press to move through the collection
B
Switch between surround collection and stereo track
A
Return to home
HOME
Zoom in or out to hear more or fewer songs
1
2

Figure 6.2: Initial user interface for the amblr.

point in the virtual space, localisation errors may cause the nearest song to appear to be elsewhere. The confirmation is needed then to make sure that the song the user believes they are selecting is in fact the nearest song. This is akin to a dialog box in a graphical user interface confirming an action.

6.1.2 User interface

The amblr's user interface consists of a Wii remote and headphones. The Wii remote has 3 accelerometers, an IR camera, 7 buttons, 4 buttons arranged in a directional cross, 4 LEDs, the ability to vibrate, and a speaker to play audio. The Wii remote is used as an inexpensive prototype of a mobile device. It contains the same sensors and controls found in many handheld devices. The mapping of the controls to the buttons on the Wii remote can be seen in Figure 6.2.

The user moves through the collection by pointing the Wii remote in the direction they would like to move while pressing the 'B' button. If the user wishes to move backwards, they need to point over their shoulders towards the desired direction as illustrated in Figure 6.3. The data from the accelerometers is processed to determine the direction that the remote is pointing in two dimensions. The user then moves with a constant velocity in the direction dictated by the remote. The remote can easily move without the user intending it to, so the 'B' button is used to indicate when a movement is intentional. The accelerometer data is only read when the button is pressed.

The user presses the '–' button to zoom out and hear more songs and the '+' button to zoom in to hear fewer songs. When the user is sufficiently close to a song, the remote vibrates indicating that the song can be heard in stereo as shown in Figure 6.4. By pushing the 'A' button, the user leaves the two-dimensional browser space with multiple songs playing and hears only the nearest song in its original stereo format. When the user is finished listening in the stereo environment, they can return to the virtual space and select another song by pushing 'A' once more.

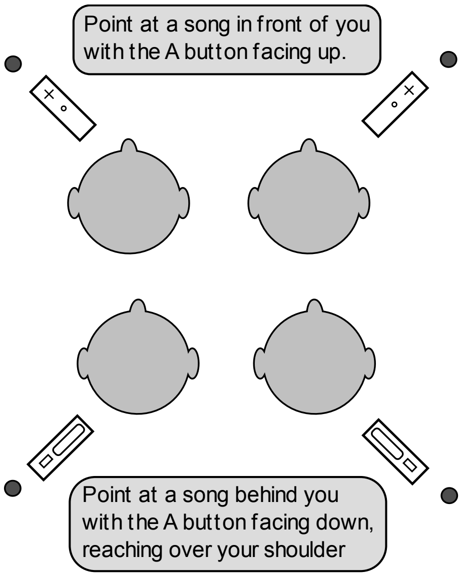

Point at a song in front of you with the A button facing up.
Point at a song behind you with the A button facing down, reaching over your shoulder

Figure 6.3: Instructions for how to access songs in front of and behind the user.

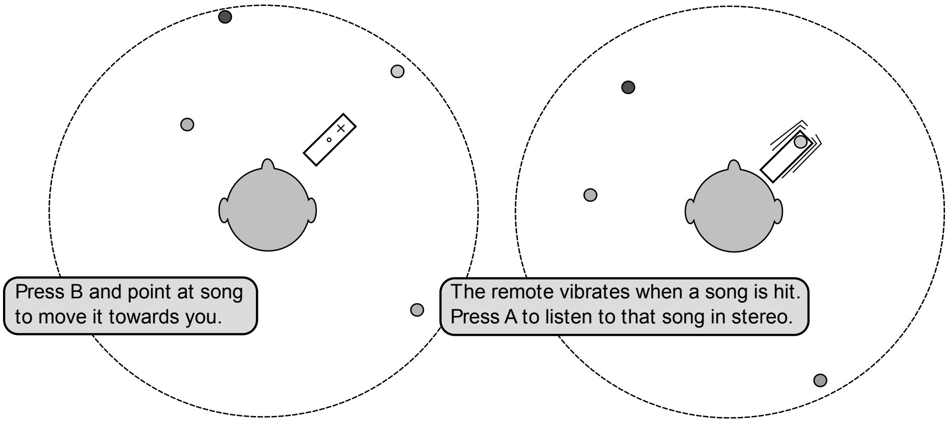

Press B and point at song to move it towards you.
The remote vibrates when a song is hit. Press A to listen to that song in stereo.

Figure 6.4: When a song is close enough to the user, the remote vibrates indicating that the song can now be heard in stereo.

The user returns to the 'home' location at the centre of the map by pressing the button labeled 'HOME' on the Wii remote.

6.1.3 Implementation

The first implementation of the amblr is built in Max/MSP,[1] a graphical programming environment for music creation and processing. The environment allows for fast prototyping and easy integration with controllers such as the Wii remote.

The Wii remote's data can be accessed through the aka.wiiremote[2] external for Max/MSP. It connects to the computer via Bluetooth. Most of the capabilities of the remote are not being used, but only using data from the accelerometers, 4 of the buttons, and the vibration mechanism. The vibration mechanism can only be turned on or off and does not have any finer control. While it can be difficult to extract precise directional

[1] `http://cycling74.com/products/maxmspjitter/`

[2] `http://www.iamas.ac.jp/~aka/max/#aka_wiiremote`

information from the accelerometers, using the more absolute measurements from the IR camera requires two IR emitters. This requires additional hardware and IR cameras are not standard in mobile devices. By using only the accelerometers there is no absolute direction that the remote needs to be pointed towards. If using the IR camera, the remote needs to be pointed towards the IR emitters. With the accelerometers the user can be facing towards or away from the computer and it has no effect on the direction of movement within the interface.

6.1.3.1 Spatial audio rendering

The audio is rendered with virtual Ambisonics. As the initial development of the amblr was conducted before the evaluation in Chapter 3 was complete, the implementation uses a less than optimum approach. However the evaluation showed that higher order virtual Ambisonics performs the same as first order virtual Ambisonics in listening tests.

The spatial audio engine uses a set of externals for Max/MSP produced by the ICST[3] for the Ambisonics encoder and decoder, encoding and decoding to third order. The horizontal information is decoded to eight loudspeaker feeds which are each convolved with the non-individualised HRTF pair for the loudspeaker's position. The virtual loudspeakers are in an on-axis configuration. The HRTFs are from the set of compact HRTFs produced by MIT [Gardner and Martain, 1994].

6.1.4 Evaluation

The system was evaluated by 12 users, 4 female and 8 male, with varying experience with the Wii or other gaming consoles. Users were given time to become familiar with the interface, usually taking about 10 to 15 minutes, and then answered 4 questions rating various aspects of the interface on a scale of 1 to 7 and 3 questions with open responses. Though the users did not have a visual interface, their movements and actions were monitored on a visual display behind the user.

The instructions and questionnaire given to each participant can be found in Appendix C.

6.1.4.1 Data set

The evaluation data set is a collection of tracks positioned in a two-dimensional space structured by mood by Levy and Sandler [2007]. Dimension reduction techniques were applied to a large corpus of tens of thousands of mood words mined from social tags for a collection of several thousand tracks to create an updated low-dimensional emotion space for music. To create a two-dimensional arrangement of tracks, the plane was defined by the first and third most significant axes, which correspond roughly to activity and valence. Each song was then mapped to the centroid of the three emotion words most frequently applied to it in the data set of social tags. The weight associated with each word was the number of times the track had been tagged with it. The set of 320 songs distributed across two dimensions is seen in Figure 6.5.

[3] `http://www.icst.net/research/downloads/Ambisonics-externals-for-maxmsp/`

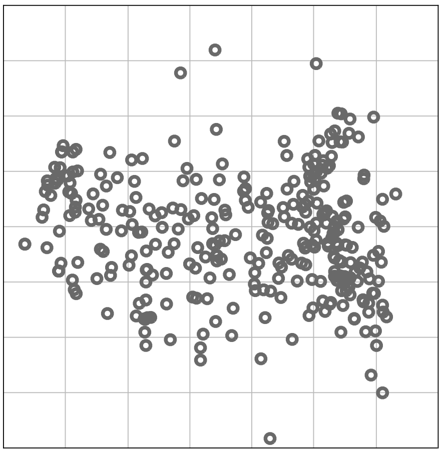


Figure 6.5: A overview of the spatial distribution of the songs for the data set used during evaluation. Each song is a circle; the scale is merely relational amongst the songs, there are no absolute units.

6.1.4.2 Results

Histograms of the responses to the four questions with rating scales can be seen in Figure 6.6. The responses to the question “how easy is it to travel through the collection” were evenly split. Six people gave a negative response of a two or three and six gave a positive response of a five or six. Those that rated the task higher seemed to learn the interface faster, but all the users found the interface confusing when there were no songs within the listening range. They also had difficulties interpreting when and where they were moving without any audio cues. Several users suggested a feedback mechanism to give a sense of the boundaries of the collection would be helpful.

Some users found the zoom functions beneficial especially when the data was sparse and no songs could be heard. A few found the zoom function confusing, but this seemed related to how individual users approached the interface. Some relied on the zoom function more than changing their position to locate new songs while others tended to move through the collection and use the zoom function sparingly. The users that used the zoom often exposed some bugs in the application that caused difficulties with the interface, so those that did not use the zoom as much tended to have a more positive experience with the interface.

Unsurprisingly, the most common complaint about localising songs in the song environment was the effects of front-back confusion. Once users were aware of this error they found they could move more easily through the environment by moving forwards or backwards to resolve localisation issues. Overall, the users did not find the environment difficult to understand and could separate the songs in space easily, but they had difficulties then expressing their intention to move to a specific place with the remote.

Users often found it difficult to approach a single song; this may stem from a number of factors. The data set is not loudness normalized so some songs may be significantly

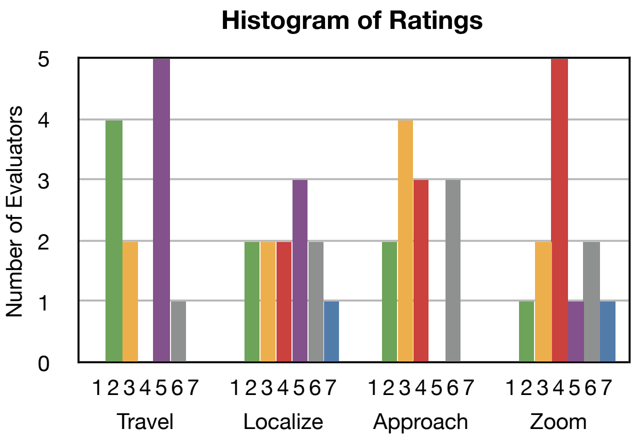


Figure 6.6: Histogram showing the distribution of user ratings for each question, “how easy it is to travel through the collection,” “how easy is it to localise a song,” “how easy is it to approach a song,” and “how useful is the zoom function.” For all questions, 1 is the most negative response with 7 being the most positive.

louder or softer than others, causing them to be perceived closer or farther than their actual coordinates in relation to other songs. Some users were more adept at moving within the world in a certain direction such as in front or to the sides, so when a song moved out of the preferred region, it was difficult to approach.

When asked about the perceived viewpoint, three users felt that the songs were static and that they were moving around the collection and nine felt that the songs were moving while they stayed in the same place. However, some users commented that they felt the difference to be ambiguous and perceived both viewpoints at different times.

The users had a wide range of responses when asked whether they would prefer a visual interface. Four felt strongly that there should be while three others responded strongly that interface should remain as audio only, noting that a visual component would inherently change the application. The other five users felt that with some improvements in the current interface, a visual component would not be needed. Most felt that the only visual information they needed was global location information, i.e. where they were located in relation to the entire data set.

There was a great deal of enthusiasm for the interface. Users felt that it has great potential and was a unique way of interacting with a collection of music, but did not think the mappings between the remote and movements in the virtual space were yet intuitive enough.

6.1.5 Discussion

Twelve users evaluated the interface and responded positively to the amblr in discussions and in the open questions on the questionnaire, but reported difficulties interacting with the data set. The evaluation confirmed that the mapping of user movement to virtual movement through the music collection is not yet ideal. This is not surprising when using the Wii remote. The remote has a large number of sensors and buttons that can be used to convey user information, in particular the three accelerometers, but this freedom in expression can be difficult to interpret and map in a universally intuitive manner. It may be advantageous to move away from the Wii remote and towards a more traditional gaming interface that is a more familiar controller for most users.

The mappings between the remote and movement through the virtual space could be arranged in a number of configurations that were not tried here. In the current configuration, users never rotate their point of view. In a manner of speaking they are always looking north when they move about the virtual space, whether forwards, backwards, or sideways. If a user could rotate the entire sound field, then they ideally could move songs to an easily accessible region, such as directly in front without changing their location within the environment. This also might ease front-back confusion. Since the audio is encoded into B-format, transformations such as rotations are easily applied [Malham, 1998].

If it is found that moving in the virtual space is not enough to overcome the errors introduced by non-individualised HRTFs, then head-tracking could be added without requiring HRTF interpolation as it would only involve rotations in the Ambisonics domain.

Since only the audio content is accessed by the user, tag-based or textual information is never used in the interface. Incorrect song title, artist, or other tags are inconsequential to the user, though if the user wishes to learn this information about the songs, then either a visual display or text-to-speech function will be needed. The content of the audio files accessed by the interface does have a strong impact on the interaction, especially with localisation. If the files are not normalized or vary greatly in volume, the songs may not be perceived in the correct locations relative to each other. This may be further exacerbated in diverse data sets where, for example, highly processed pop music may be heard near more dynamic, less compressed classical string music.

By using a virtual Ambisonics approach, much of the computational burden associated with moving sound sources using HRTFs is eliminated. While the common problem of front-back confusion when using non-individualised HRTFs still exists, the ability of a user to move and change the spatial cues presented to them helps alleviate problems. In order to simplify the interaction, the user cannot rotate the sound field, but rotations may further help reduce localisation errors. A third order Ambisonics encoder is used in this implementation of the amblr because it was the highest order the encoder could handle and the computer had no problems rendering the audio. However, as has been shown in Chapter 3, such a high order is not necessary.

6.2 Second Design Iteration

The results from the user evaluations and further informal evaluations from demonstrations of the amblr led to design improvements for the user interface. Software engineering considerations also influenced the development as the system was moved from the Max/MSP programming environment to a server-client architecture.

This section presents the improvements made to the amblr in a second design iteration. A second formal evaluation was not conducted, though a usability study to further evaluate the amblr is proposed in Chapter 7.

The design criteria for the second iteration are similar to the first iteration with some additions and deletions. The new criteria are:

- Minimal visual display, but interaction possible without any visual display
- Allow finer control of how much information is presented
- Minimal instruction needed to use interface
- Allow rotations in the space

The interface should continue meet the following criteria:

- No hardware beyond headphones and mobile device
- Gestural control and haptic feedback using buttons and sensors available on mobile device
- No head tracking, let user move in virtual space to resolve front-back confusion
- Two-dimensional virtual space populated by music that is not restricted to any particular space
- Binaural audio rendered with virtual Ambisonics

Though the user evaluation indicated that the interface would benefit from reverberation, it is not implemented in the second iteration. This is to keep the implementation simple, but it is still an area that should be pursued in future research.

6.2.1 Interface design

The first design of the amblr interacted with a two-dimensional arrangement of songs, but did so with only gestural input and without any visual display. Several of the design features were found to be useful while others left room for improvement. The successful features from the previous interface that have been kept include a 'home' function, zooming, haptic feedback, and gestural control. The user study showed that a visual display would be beneficial to some users and that the gestural control needed some refinement. Additional considerations include exaggerating focus and controlling the playback of concurrent songs through cyclical playback.

6.2.1.1 Navigating the map

As was established in the first design iteration, the primary user interaction with the map of music is moving around the two-dimensional space towards and away from various songs. This is done in a similar manner to first person viewpoints in video games. The user occupies a point on the map and surrounding songs are spatialised according to their location and proximity to the user as defined by the map. This is illustrated in Figure 6.7. The same map is shown with four different sets of user interface parameters. For all four maps the listening position is identical and is at the center of the concentric circles.

In the first design iteration, the user could not rotate the sound field, but only move forward and turn left or right, but not move backwards or 'step' to the side. The second design iteration uses an alternative approach: the user can only move forward and rotate the sound field, but not move backwards or 'step' to the side. The analogy of driving a car which cannot go in reverse is used. As will be seen in Section 6.2.2 this analogy is extended to the physical controller. As was discussed in Chapter 3, rotating in place is computationally efficient with virtual Ambisonics.

To further exaggerate the audio environment, sound sources to the front of the user are louder than sources the same distance away but to the sides or rear of the user. This is to emphasise the information from the front. Since no visual information is available to identify objects of interest the audio compensates. This is implemented as a user-adjustable *focus of perception* in soniXplorer [Lübbers and Jarke, 2009]. The idea of mimicking visual focus is also explored in the AudioStreamer [Schmandt and Mullins, 1995] and Audio Hallway [Schmandt, 1998] where the volume of sound sources increases by 10 dB when the user turns towards a source to indicate interest.

In informal user tests with the amblr, the focus was adjustable, as it is for soniXplorer [Lübbers and Jarke, 2009]. However, it was found that users were being overwhelmed with too many parameters, so the focus is fixed.

The gain of a song is inversely proportional to its distance:

$$g_{distance}(d) = min(1, \frac{r}{d} - \frac{r}{d_{min}}) \tag{6.1}$$

where the gain $g_{distance}$ of a song is dependent on its distance d, the rate of decay per unit r, and the minimum distance d_{min} from the user where the sound source begins to decay.

Songs directly in front of the user play at a volume only attenuated by the distance, but sources to the sides and rear have additional attenuation. A song located at φ radians relative to the direction the user is facing has a gain of:

$$g_{focus}(\varphi) = e^{-\varphi^2/4} \tag{6.2}$$

The gain for a song s located at position $\vec{p_s}$ with the user at position $\vec{p}$ is then the product of the attenuation from the distance and the angle of the song relative to the user:

$$g(s, \vec{p}) = g_{distance}(||\vec{p} - \vec{p_s}||) * g_{focus}(\angle(\vec{v}, \vec{p_s} - \vec{p})) \tag{6.3}$$

where $\vec{v}$ is the user's orientation.

Front-back confusion occurs with non-individualised HRTFs and was found in the initial design of the amblr. The introduction of focus aims to lessen this error by attenuating sound sources that are not directly in front of the listener.

6.2.1.2 Zoom

The zoom functionality is the same as described in the first design iteration. The only alteration is that the maximum area the user can listen to corresponds with the distance gain attenuation described above. The interface allows the user to zoom out only until the distance of a sound source directly in front of the listener has a gain of zero. A larger listening area would then have no effect as all sources would be attenuated due to distance. The zooming function is illustrated in Figure 6.7.

6.2.1.3 Auditory Landmarks

The auditory landmarks in the first design iteration were dependent on the data set used in the evaluation. The data set consisted of music clips ranging from 30 to 90 seconds in length. Though the second design iteration does not use a specific data set, all songs are limited to 20 seconds. A length of about 30 seconds is commonly used in the literature, but it was felt that this is slightly too long for an auditory landmark.

The 'home' function is identical to the first design iteration – when requested, the user is returned to the centre of the map. Other parameters, such as the orientation and zoom level, are not changed.

6.2.1.4 Selection

The method for selecting a song from the collection was found to be successful in the first design iteration as users were able to select individual songs and was not altered in the second design iteration. A user can indicate that they would like to select the nearest song by pressing a button on the physical controller. The selected song is then the only song heard outside of the virtual space to confirm that it is the desired selection.

6.2.1.5 Cyclical playback

The cyclical interface described in [Ali and Aarabi, 2008] mixes multiple audio files into a single audio stream, weighting each individual audio file to reflect its ranking as a result to a query (no spatialisation is used). The effect is a short song clip fading out as another clip fades in so that multiple songs can be quickly reviewed. Visualisation of the audio provides metadata and helps the user to differentiate between songs. The audio signal is analysed in order to automatically present the sections of the song that may be of the most interest. The sections with the highest level of information in the frequency domain, high spectral entropy, are considered the most interesting. Only the first 20 to 30 seconds are considered.

Presenting multiple songs simultaneously can become overwhelming to some users, so the amblr uses a similar cyclical presentation which lets the user control how many concurrent song can play. When a user decides to listen to only a single song at a time, they can still listen to multiple songs over a shorter period of time than would be required if they listened to each song sequentially in an interface such as iTunes. A clip of a song is played and then is immediately followed by a clip of another song. The songs played are all of the songs within the listening area defined by the zoom function. Each clip is played from the location defined by the map and relative to the user's listening position.

All songs within the listening area are heard within 20 seconds of entering that area. In the top left map in Figure 6.7, the largest circle or most zoomed out listening area contains 17 songs. A 1.17 second clip of each of the 17 songs would be played before repeating any of the songs. If the user zooms in to the smallest listening area which

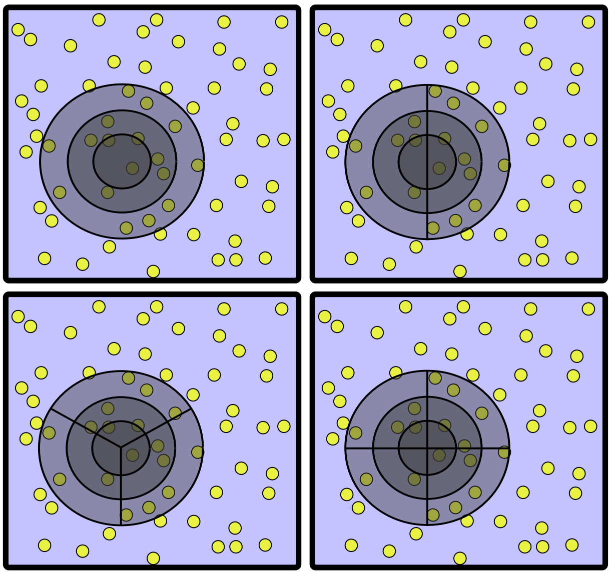


Figure 6.7: Each square is an illustration of the same map with different interface parameters. The listening position is identical in each map and at the center of the concentric circles. The zoom level increases as the diameter of the circles in each map decrease so that less of the total map and only the nearest songs are heard. The number of concurrent sources allowed increases in each map.

contains 2 songs, each song would be played for 10 seconds.

The length of time a song is played is:

$$t(n) = \begin{cases} 20/n & \text{if } 1 < n < 66 \\ 0.3 & \text{if } n \geq 66 \end{cases} \tag{6.4}$$

where $t(n)$ is the time in seconds that each song is played when there are n songs.

The user can choose to listen to up to four songs playing concurrently. Concurrently playing songs are best distinguished if they are in different locations. If more than one song is allowed to play at the same time, the listening area is dividing into regions to encourage concurrent playback of songs from differing locations. The length of time each song is played is then determined by the number of songs in that portion of the playback region. For example, in the top right map in Figure 6.7, the listening area is divided into two portions so two songs at a time will play concurrently. In the largest, zoomed out listening area, 11 songs will play for a little under 2 seconds each to one side of the listener and 6 songs will play for a little over 3 seconds each to the other side.

6.2.2 User interface

The amblr has two modes of input: a physical controller (a Wii remote within a steering wheel holder, see Figure 6.8) and a graphical user interface (GUI) with a mouse. The visual display is optional if using the physical controller instead of the mouse. The audio output is identical and all functionality is completely duplicated when using either controller.

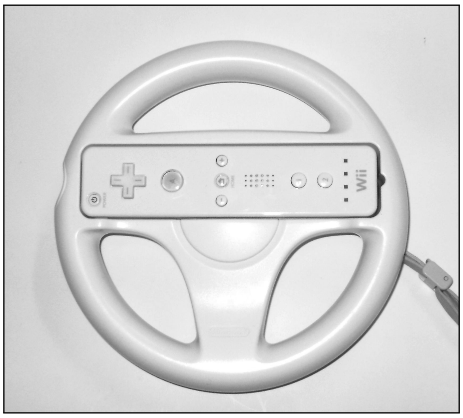


Figure 6.8: Wii remote within a steering wheel holder commonly used for racing games.

6.2.3 Graphical user interface

The first design iteration of the amblr used a Wii remote to navigate a two-dimensional map of songs without any visual display, but this was found to be confusing to users. They were easily disoriented in the virtual space without any visual cues, problems which were further compounded by front-back confusion with the binaural audio.

The amblr now has a visual display that can be seen in Figure 6.9. It displays the location and orientation of the user within the two-dimensional map to give the user more structure and prevent them from becoming lost. The interface explicitly does not show the location of any of the songs nor does it display any metadata about the collection. This is so the auditory display is still the primary mode of interaction and is only complemented, but cannot be replaced, by the visual display.

Various graphics on the visual display illustrate the current values of the parameters of the interface. Directly clicking on those graphics alters the values. The illustration of the user's location and orientation within the map is the only graphic that cannot be directly manipulated.

The central section of the visual display is a series of concentric circles. The outermost grey circle and innermost white circle do not change their sizes, but the middle purple circle will change its diameter. The new diameter adjusts to where the user clicks the mouse with the white and grey circles being the minimum and maximum sizes. The size of the purple circle controls the zoom level with the larger the circle the larger the listening area of the map.

Clicking the white circle in the center of the concentric circles selects the nearest song. Only that song is then heard; if any other songs are playing, they are muted. Clicking on the white circle again brings back the full virtual space and all songs are heard according to the parameters of the application.

The collection of graphical buttons at the top of the concentric circles allow the user to move around the virtual space. The triangle pointing up moves the user forward and the other two arrows rotate the user to the left or right 5°. By clicking on the icon of a house, the user is returned to the home position at the centre of the map.

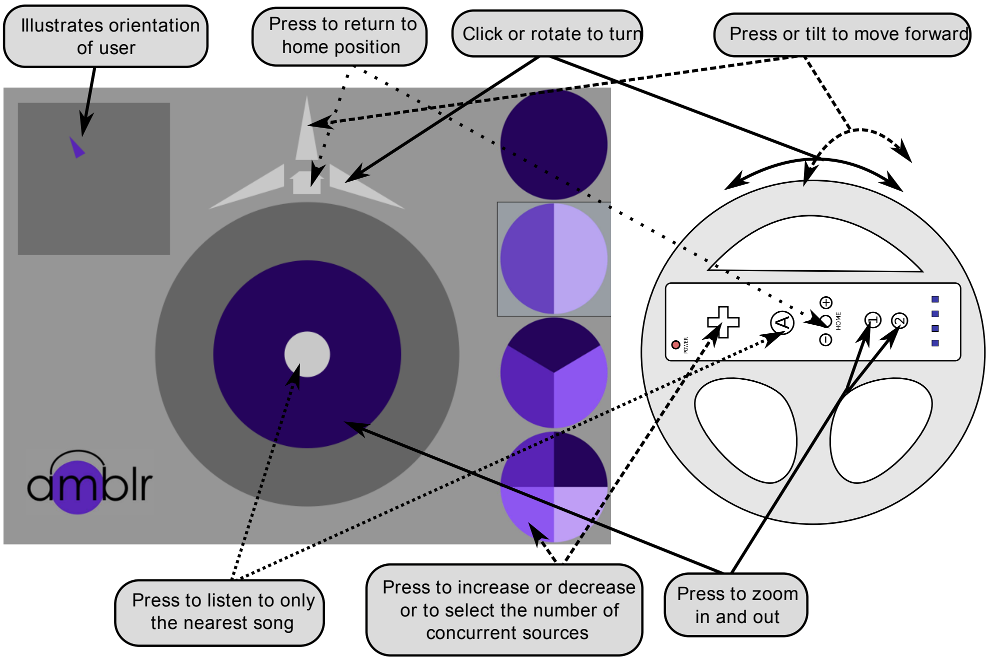

Illustrates orientation of user
Press to return to home position
Click or rotate to turn
Press or tilt to move forward
amblr
Press to listen to only the nearest song
Press to increase or decrease or to select the number of concurrent sources
Press to zoom in and out

Figure 6.9: The GUI and diagram explaining how the mouse and Wii remote are used to control the amblr.

On the right-hand side of the visual display are four circles each divided into one to four partitions. They each represent the number of concurrent sound sources and illustrate how the listening area is partitioned. The user clicks on the appropriate circle to change the number of sources.

6.2.3.1 Haptic interface

The initial design of the amblr used a Wii remote for gestural interaction, but the gestures used were not intuitive enough for users. The user held the Wii remote and pointed towards the intended direction to move, but had to take care when trying to move backwards. Most users had difficulties navigating the virtual space. The interface after the second design iteration also uses a Wii remote, but it is held in a steering wheel configuration and holder as for Wii racing games, seen in Figures 6.8 and 6.9. The steering wheel gives context for how the physical interface is used and is less ambiguous as to how it is held.

The accelerometers are used for rotating the sound field and for moving the user forward, identical to the forward, left, and right arrows in the GUI. Turning the steering wheel to the left or right as if driving rotates the user in the corresponding direction. Tilting the steering wheel forward moves the user forward. The rest of the parameters are controlled by buttons.

The Wii remote has a cross button on the left of the controller, a large circular button in the center, and two smaller circular buttons on the right. The user increases the number of concurrent sound sources by pressing either up or right on the cross button and can decrease by pressing down or left. The zoom level is increased by pressing the left smaller circular button (labeled ‘2’) and decreased by pressing the right smaller circular button (labeled '1').

In the center of the Wii remote is a button labeled 'HOME.' Pressing the button will return the user to the centre of the map.

The Wii remote vibrates when the user approaches and 'hits' a song. If the user presses the large circular button (labeled 'A'), they hear only the nearest song. If they hit the button again they return to the full virtual environment. The user can listen to the nearest song without the Wii remote vibrating, but the vibration helps the user confirm that they have approached a song.

6.2.4 Implementation

Max/MSP was found to be a limiting environment for development. It does not handle dynamic memory allocation effectively and does not communicate via Bluetooth with a Wii remote reliably. It also is a proprietary programming environment which limits its portability to desktop computers, not mobile devices. A decision was made to use freely available libraries with the exception of the application used to communicate with the Wii remote, Osculator[4]. It is chosen because it is easily integrated in the user interface and offers valuable tools for interpreting accelerometer data.

The amblr system is based on a client-server architecture as illustrated in Figure 6.10. While the entire system is run locally on a single computer, the design is intentionally kept flexible so that it could instead stream audio from a central database of music to a client running the system.

The backend of the system is run by a CherryPy[5] server. The audio is rendered with custom Python libraries and uses Python bindings to PortAudio[6] to output binaural audio to the soundcard. The map of songs is managed by a PostGIS[7] database. PostGIS is a spatially-enabled PostgreSQL[8] database. The database is queried with SQL to determine what songs are nearest to the user's position on the map.

The Wii remote connects to the computer through Bluetooth. Osculator is used to map the accelerometers, buttons, and haptic feedback to OSC (Open Sound Control)[9].

The GUI is written in Java using Processing[10] to render the visual display and handle mouse input. The Java client also sends and receives OSC to the Wii remote via Osculator. The client updates the Python server with HTTP requests using a REST-inspired structure. It deviates from REST because a client-stateless-server architecture is not maintained [Fielding, 2000]. Instead, the client is considered stateless while the server maintains the parameter values and context until updated with new parameters by the client. The server is updated with the user's location, orientation, zoom, and number of concurrent sound sources. Information such as the user's current location and orientation and whether the user has 'hit' a song is reported to the client

[4] `http://osculator.net`

[5] `http://www.cherrypy.org`

[6] `http://www.portaudio.com/`

[7] `http://postgis.refractions.net/`

[8] `http://www.postgresql.org/`

[9] `http://opensoundcontrol.org/`

[10] `http://processing.org`

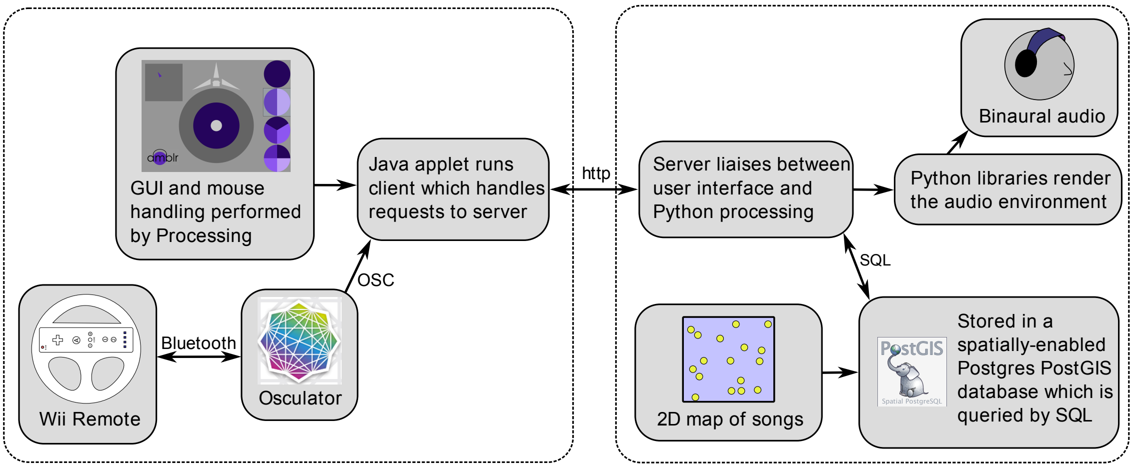

amblr
GUI and mouse handling performed by Processing
Java applet runs client which handles requests to server
http
Server liaises between user interface and Python processing
Python libraries render the audio environment
Binaural audio
OSC
SQL
Wii Remote
Bluetooth
Osculator
2D map of songs
PostGIS
Spatial PostgreSQL
Stored in a spatially-enabled Postgres PostGIS database which is queried by SQL

Figure 6.10: System architecture for the second design iteration of the amblr.

via JSON.[11]

6.2.4.1 Spatial audio rendering

The audio is rendered using the virtual Ambisonics algorithm described in Chapter 3. First order decoding is used with on-axis virtual loudspeakers.

6.2.5 Evaluation

No formal evaluation was conducted, but approximately 40 individuals have given feedback throughout a 9 month period. The users have had brief introductions to interface and have tried using the interface for approximately five minutes. The GUI has helped clarify how the use the interface and the intended function. However, the large number of controls have made it difficult to succinctly instruct a new user as to how to interact with the interface.

6.2.6 Discussion

The amblr is a novel music browser that relies on an auditory display as its primary mode of interaction. It brings together successful elements of previous designs that had not before been compiled into the same interface while introducing further novel design features.

The evaluation in the first design iteration showed that improvements were needed, but that some portions of the previous interface did work well. The amblr improves upon the previous design by using an optional visual display, developing a different approach as to how the Wii remote is used, and by further reducing the number of concurrent songs by introducing a cyclical approach. The second design iteration kept the 'home' function, selection method, and haptic feedback developed in the first design iteration.

Future work will look towards further design improvements including a global view function and auditory icons. Users worried that they were beyond the edge of the map and moving away from the songs. A single auditory message such as a short tone could tell them when they've reached the edge of the space. Additionally, it was suggested by a

[11] http://www.json.org/

user of the second iteration during an informal evaluation that a continuous background sound could be played. Then even if no songs are within range to be heard, the user still knows that the system is functioning.

While the zoom feature allows for a broader overview of a larger portion of the map, it can still be difficult to gain insight from a global perspective. One potential solution is the approach taken in the city maps interface described in [Heuten et al., 2007]. A user can step back from the first person exploration of the auditory space and explore the same two-dimensional map projected onto an auditory space in front of the user. The interaction is then similar to the SoundTorch [Heise et al., 2009].

As discussed in Section 3.1, auditory icons are a major portion of auditory display design for tasks other than exploring music, none of the interfaces cited here nor the amblr make use of them. The amblr could benefit from auditory icons, especially if it is used without the visual display. Frustration with an interface occurs when it no longer appears to be responding. The user may assume that the application is no longer functioning, when instead they are misunderstanding how to use it or are asking it to do something incorrectly. In the previous user study, users expressed frustration and concern when they did not hear any audio. Users believed that they have gone beyond the edge of map but they may have merely been in an area on the map without any songs. Auditory icons may be a way to resolve this issue.

The best spatialisation approach was chosen given the system requirements. Virtual Ambisonics was found to be favourable over convolution with HRTFs in its flexibility and low computational cost. The client-server architecture implemented is not directly scalable to an online interface as the audio is rendered and directly sent to the sound card. However, audio streaming could be implemented so that server and client would not need to exist in the same location, as shown by Mariette et al. [2010]. Virtual Ambisonics is particularly advantageous for a client-server architecture as B-format audio can be streamed and then decoded to a binaural signal client-side. By decoding within the client, the audio signal can essentially be cached to reduce delay when rotating. Rotations from the interface or head-tracking, if used, can be applied without needing to update the server.

Further evaluation is needed to determine the effectiveness of the design including how it compares with standard tools for browsing a collection of music without audio as the primary mode of interaction. The user evaluation study with 12 participants indicated that the new design tools implemented in the amblr will improve user interactions, though another user study is necessary to confirm this.

6.3 Art Installation

A virtual two-dimensional map of music can be a difficult concept for some users, particularly without a graphical representation of the space. It is essential that interaction with the map is intuitive and does not hinder the user from accessing the audio content, but instead enhances the collection. Instant usability is also important as the user may not be willing to expend much energy learning a new interface.

A literal interpretation of moving around a map of music was explored in a reinterpretation of the amblr as an art installation. The map of music is no longer occupies only virtual space, but physical space. The map is projected onto the floor allowing a user to physically move around the map by walking instead of relying on a controller. This removes the learning curve and gives instant usability to the interface; users can quickly start exploring the music collection instead of spending time learning how to explore the music collection.

The amblr uses the *museum metaphor* [Cohen, 1991] to navigate a collection of music. The sound sources are static and the listener moves around the virtual space. The installation implements the *cocktail party metaphor* as the sound source and listeners move around the space. The listener is now also a sound source, generating sound and no longer being a silent voyeur of the virtual space.

The physical interface was installed as an art piece in the Information Aesthetics Showcase at the Association for Computing Machinery SIGGRAPH Conference in 2009. The conference attendees consisted of professional artists, designers, and engineers from industry and academia. Members of the general public also attended, often as guests of the conference attendees. The conference is not primarily focused on audio interaction, so most of the users were unfamiliar with the technology and binaural audio.

The installation allowed users to explore a collection of music relevant to the history of the physical location of the conference, New Orleans, Louisiana. While exploring historical audio recordings, the user's exploration was influenced by the recent history of past users of the interface. The music from the collection was not assigned a static location on a map, but was mapped to the movement of a previous user from when they explored the map. The songs from the collection then moved around the map often causing the user to chase a song if they wished to continue listening to it. The current user's presence in the space also influenced future users as their movements were mapped to another song from the collection.

The art installation was inspired by the amblr, but did not replicate all of its functions in order to simplify the interface. In general, all of the parametric controls were removed. The user could move around the map, but there was no zoom function, exaggerated focus, home position, or cyclical playback. Concurrent songs were presented to the user and they could influence their loudness by moving in the physical space. The user wore closed-ear headphones and could not easily hear any sounds outside the interface. They were aurally separated from the physical room around them.

6.3.1 User interface

The user enters the space wearing wireless headphones and is free to walk around a 3 m by 4 m space. The space is defined by a series of images projected onto the floor. Upon entering, the user is in the centre of circle. Other identical circles are also in the space and are moving around.

The user immediately hears music upon entering the space, though it may be distant. As many as three or four songs may be playing and moving around the space. Each of

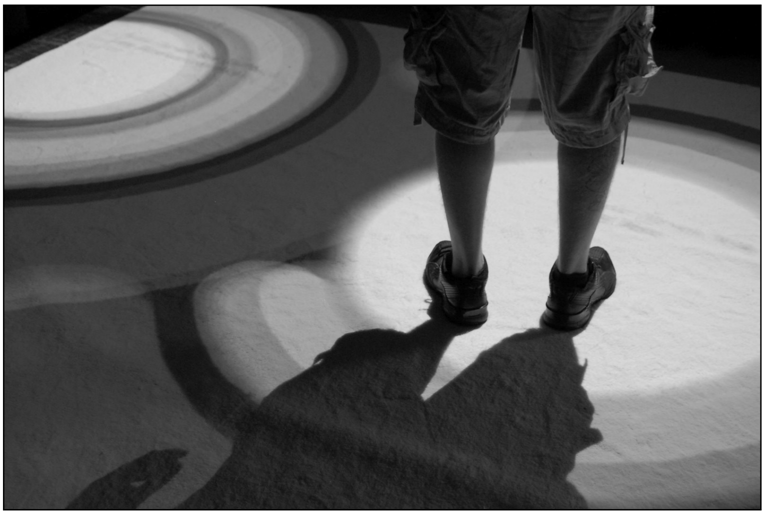


Figure 6.11: A photograph of the installation in use.

the moving circles projected onto the floor is a moving song as seen in Figure 6.11. The metadata associated with that song appears on the circle.

The songs in the space are not mapped to static locations, but are mapped to the locations of previous users of the interface. When a user enters the space, their movements are recorded. A song is then assigned to the path that the previous user took. A random sampling of songs with an assigned path are chosen and populate the virtual space. The next user to enter the space then hears those songs and can observe their movements. At the same time, the current user's movements are being recorded and will be recalled in the future with another user.

6.3.2 Implementation

The physical interface is very similar to the server-client architecture in the second design iteration and is illustrated in Figure 6.12. A CherryPy server manages all input and output. The amblr's audio rendering engine generates a binaural signal with virtual Ambisonics. However, only distance is used to attenuate a sound source; there is not an exaggerated focus function.

The binaural audio is sent wirelessly to Bluetooth headphones. A webcam tracks the head rotation and position of the user in the space by following three infrared LEDs mounted on top of the headphones. The webcam has a filter placed over the lens so only infrared light is received. The three LEDs are then the only items within the view of the camera. The image is transformed to black and white and tracked using optical flow with the OpenCV library[12].

A projector pointed at the floor highlights where the listener is positioned and the positions of the other tracks in the space. The graphics projected onto the floor are driven by a Flash script which is updated by the CherryPy server.

[12] http://opencv.willowgarage.com/

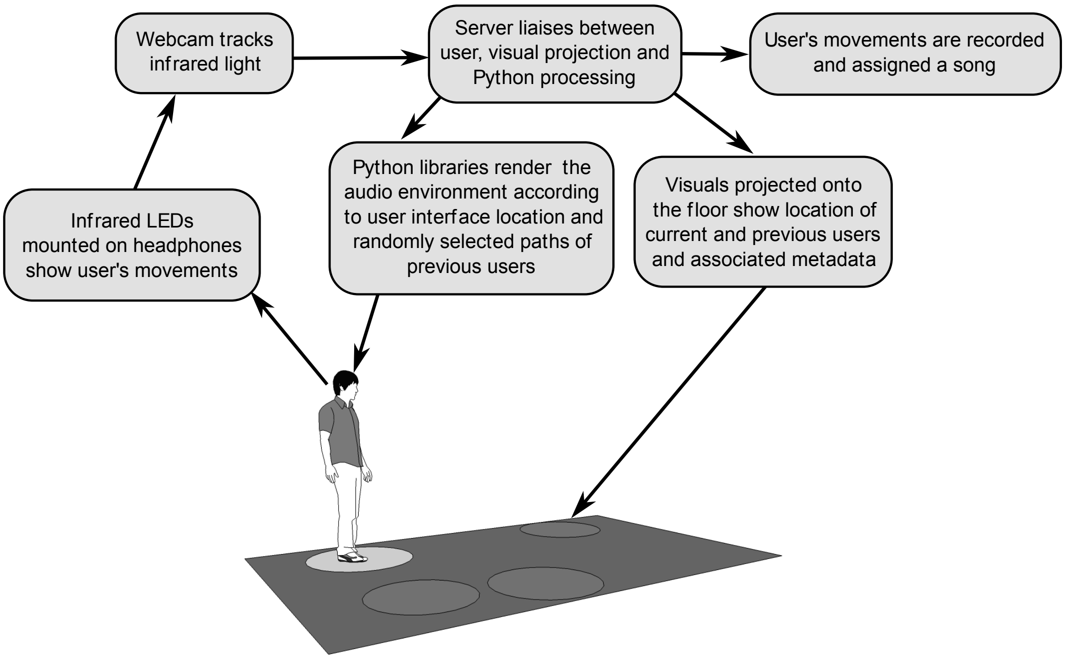

Webcam tracks infrared light
Server liaises between user, visual projection and Python processing
User's movements are recorded and assigned a song
Python libraries render the audio environment according to user interface location and randomly selected paths of previous users
Visuals projected onto the floor show location of current and previous users and associated metadata
Infrared LEDs mounted on headphones show user's movements

Figure 6.12: Implementation of the art installation.

6.3.3 Data set

The installation was interested in exploring a collection of music that would not be familiar to the general public, but held some educational or informative value. As the installation was in New Orleans, Louisiana, an ethnomusicological collection of music from the southern United States was selected. The collection consisted of 35 recordings from Alan Lomax's training tapes [Lomax, 1976]. Alan Lomax was a field recorder and ethnomusiologist who traveled the world recording folk music.

6.3.4 Discussion

Approximately 1000 users explored the interface over five days. They ranged from young children to students to adults. Overall the response was very positive. There was small segment of the population, around 5%, that was confused by the interface. They tended to be older and were usually resistant to listening to the interface.

The amblr was used an inspiration and technical foundation for an art installation which allowed members of the public to explored a collection of historical recordings. As for the amblr interface, the music was arranged on a two-dimensional map, but the participants did not need to use a controller such as a Wii remote to navigate the map. The virtual space was mapped onto a physical space and participants needed only to walk around the physical space to explore the collection. Controls such as zoom and cyclical playback were not available, but further interaction with the collection was facilitated through the mappings of the music. The music in the collection was not static, but moved according to where past participants walked while in the space.

The majority of reactions to the installation were positive, and the installation had universal appeal to people with varying ages and different combinations of technical and artistic backgrounds. Perhaps one of the main testaments to its success was that people that had explored the installation often returned with friends recommending them to try it out.

People enjoyed the experience of approaching and listening to a virtual object through physical activity without a controller mediating the movement. The novelty of the physical interface was perhaps more appealing to many people than the content of the music collection. While the concept of sources moving according to past users was sometimes too abstract for everyone to grasp, universally users were able to quickly learn how to use the interface. Their actions within the space resulted in direct results both visually and aurally.

Given further technical development, the installation could be extended to allow more than a single user to interact with the music space at the same time. Many users expressed an interest in such an interface. Ideally a future installation would map songs onto 'live' users so that multiple people could explore the space at the same time and explore the collection of music by approaching each other.

6.4 Summary

This chapter presented the amblr, an interface to explore a collection of music by allowing a user to navigate a collection without relying on text. The driving concept behind the amblr is that for effective music exploration and discovery, audio content should be presented before metadata. The amblr does this by building upon previous interfaces that use spatial arrangements of audio files, with the Sonic Browser [Fernström and Brazil, 2001] and the cyclical interface described in [Ali and Aarabi, 2008] being the primary influences.

The amblr navigates a two-dimensional arrangement of music, a popular approach used in a number of interfaces. Yet even with its popularity, it still can be improved upon. All of the previous interfaces that use a similar approach to navigate a collection of music rely heavily on visual displays to accompany auditory displays. The amblr has a visual display, but it is minimal; the auditory display is the primary mode of interaction. Although the interface is fully functional without any visuals, the GUI does aid interaction by visually illustrating the different parameters of the interface.

The amblr is designed for mobile devices as music is increasingly being consumed in a mobile context. Mobile devices are commonly used to listen to a music collection and also to add to a music collection through online purchasing or streaming services. Auditory displays are particularly beneficial for mobile devices as they have small screens that cannot present large or complex visual displays. The amblr takes further advantage of computing on handheld devices by using gestural controls and haptic feedback driven by integrated sensors such as accelerometers.

The initial design of the amblr sought to select a song from a collection of music with no previous knowledge about how the collection is arranged, only that it is in two-dimensional virtual space. It also had no visual information and was completely driven by audio and gestural control with a physical controller. Virtual Ambisonics was determined to be the most appropriate audio spatialisation algorithm for the auditory display.

A user study with 12 participants evaluated the first design iteration with the findings of the study directly informing the second design iteration. Improvements include a visual display which is complementary to the auditory display and a different set of gestures to control the interface. Mouse control with a graphical user interface was introduced as a traditional controller that does not need additional instruction in order to use. The second design iteration also led to the development of cyclical playback of content in order to reduce the number of concurrently playing songs while still quickly exploring multiple songs.

Additional user evaluations need to be conducted to conclude if the changes since the first user study lead to better interaction. An evaluation should study how users perform the same task using the amblr and also using a standard audio collection browser such as iTunes. It has been shown that an auditory display without a visual display can be equally as effective at a playlist-building task as the same auditory display with a visual display [Pauws et al., 2000]. Users just tend to be slower without a visual display, but no less accurate at using the interface. However, users have also been shown to spend less time deciding which of two playlists they prefer when they do not have any metadata presented and can only listen to the content [Barrington et al., 2009]. An auditory display such as the amblr may then prove to be as effective as or even better than a similar music collection exploration tool without an auditory display, and perhaps even more efficient as well.

A usability study will also identify additional design features that are currently missing and help refine the ones already in place. For example, the amblr does not make use of auditory icons, but the user evaluation that has been conducted found that users are uncomfortable with silence. A constant feedback mechanism is required; if there is not a visual indicator of a song, then an aural one is needed. A more immersive and realistic environment may also be simulated with additional spatialisation cues. The spatialisation algorithm does not include any reverberation; distance is only indicated through attenuation of the sound source. Externalisation could then be significantly improved by implementing reverberation, if only early reflections [Begault et al., 2001].

Cunningham et al. [2003] discuss that interfaces for music collection exploration need to include more social interaction as traditional ways of finding new music in physical stores is often a social activity. Both the amblr and the installation could potentially incorporate a social aspect by allowing multiple users to explore a collection together.

Both the user study and the installation elicited positive reactions, though with caveats that there could be further improvements. While the user study showed that the interface is not yet intuitive, the installation demonstrated that the underlying concept of navigating a two-dimensional map of music is useful if the interface does not impair interaction. The installation was limited to a small music collection and with basic navigation tools. The amblr allows a user more refined control in how they interact with a music collection, but it lacks the instant usability and ease of use of the installation. The amblr should strive to be as engaging as the installation.

Chapter 7

Conclusions and Future Work

This chapter draws together the work described in the previous chapters. Each chapter answered several research questions, but also identified new ones. This thesis concludes by expanding on those questions and outlining further work.

7.1 Summary

This thesis set out to answer *"How can audio be better incorporated into how people access information?"* This was investigated through three sub-questions:

1. Mobile applications have processor and memory requirements that restrict the number of concurrent static or moving sound sources that can be rendered with binaural audio. Is there a more efficient approach that is as perceptually accurate as the traditional method?

2. Complex acoustics models require significant amounts of memory and processing. If the memory and processor loads for a model are too large for a particular device, that model cannot be interactive in real-time. What steps can be taken to allow a complex room model to be interactive by using less memory and decreasing the computational load?

3. Commercial interfaces for music search and discovery use little aural information even though the information being sought is audio. How can audio be used in interfaces for music search and discovery?

While investigating each question, the requirements of mobility, flexibility, and accuracy have been retained. It can be concluded that virtual Ambisonics is an efficient and perceptually accurate method for creating a spatial audio environment with noise signals placed on the horizontal plane without head tracking. Complex acoustics models can be interactive when a common mixing time is found with the introduced metric based on kurtosis. The late portions of a collection of IRs can then be spatially averaged and incorporated into an interactive room model which reduces memory and processor loads. Virtual Ambisonics has also been applied to spatial auditory display for music discovery applications. A new interface has been developed which builds on previous interfaces described in the literature and also incorporate new design features. User studies have evaluated the interface and informed subsequent designs.

Chapter 2 presents the basics of human spatial hearing. Differences in time, loudness, and frequency provide the basic cues for localising a sound source, but the human auditory system has a finite resolution. How accurately a sound source can be localised depends on the location and signal characteristics of the sound source. These cues are encoded into models of acoustic spaces and those spaces can be rendered over loudspeakers or headphones.

Chapter 3 introduces the concept of spatial auditory display as a means to present information. The common design approaches are reviewed and the signal processing requirements are compiled. Virtual Ambisonics is found to be an efficient method for creating binaural auditory displays. However, it can only be adequate if it can be proven to be as accurate at rendering a spatial audio virtual as the standard method – binaural audio generated from measured and interpolated HRTFs. While virtual Ambisonics had been often used in auralization applications, it had never been fully evaluated. The effect of the decoding order and virtual loudspeaker configuration was not discussed in the literature.

Virtual Ambisonics is evaluated by analysing the ITD, ILD, and spectral errors incurred when comparing HRTFs rendered with virtual Ambisonics to the original HRTFs. Errors are observable, but listening tests show that they did not affect the perceived location of a virtual sound source. It is found that a first order decoder performs as well as convolution with HRTFs. While no difference is observed between the two different virtual loudspeaker configurations in the listening tests, the ITD and frequency response of the off-axis configuration are more accurate than the on-axis configuration for first order decoding. It can then be recommended to use virtual Ambisonics with first order decoding and an off-axis virtual loudspeaker configuration.

The evaluation in Chapter 3 concludes that virtual Ambisonics is an efficient and accurate means to render a binaural auditory display consisting of noise signals placed on the horizontal plane without head tracking. Virtual Ambisonics is then more efficient than convolution of HRTFs if more than two sound sources are concurrently rendered or if movement of the sources or head tracking is implemented.

Complex acoustics models can create realistic auralizations of historical spaces, but these models are often not interactive. Chapter 4 presents a new reverberation model based on hybrid reverberation which uses a collection of B-format IRs. A new metric for determining the mixing time of a room is developed and interpolation between early reflections is investigated. Though hybrid reverberation typically uses a recursive filter such as a FDN for the late reverberation, an average late reverberation tail is instead synthesised for convolution reverberation.

Evaluation of the mixing time metric found that it can determine the mixing time of multiple IRs from the same room. However, it is not robust enough to determine the mixing time of an arbitrary IR. Evaluation of linear interpolation between early reflections found that the error is decreased when the closest pair of IRs to the position being interpolated is chosen. Both IRs should also be approximately equidistant from the sound source. The evaluation of the late reverberation averaging found that a measured

IR can be accurately synthesised to a within an RT_{30} of 0.3 ms and C_{50} of 0.6 dB. The averaged reverberation tail is more accurate at higher than lower frequencies. This shows that this approach, while not a perfect replication of the full model, can be used to simulate a complex, non-interactive model.

Chapters 5 and 6 turned to a different application area for binaural auditory display – music discovery. Auditory display has been used in a number of interfaces for navigating audio content, but they had never before been compiled and reviewed. Chapter 5 looks at 20 interfaces and determines that several themes emerge from past interfaces. These include using a two or three-dimensional space to explore a music collection, allowing concurrent playback of multiple sources, and tools such as auras to control how much information is presented. It was found that many interfaces lack thorough testing, though there are some examples of rigorous evaluation. In general it was concluded that future interfaces should look towards social interaction and also combine elements from disparate designs to improve upon past work.

Chapter 6 built on the findings from Chapter 5, though was not able to accommodate all of the recommendations. Social interaction and more complex acoustic models within music discovery interfaces are not explored here. The amblr is a binaural auditory display for music discovery which is developed in two iterations. It was developed because virtual two-dimensional spaces populated by music have been a common approach, but not yet a perfected one. The interface is intended for mobile devices, so it uses audio and gestural control as the primary modes of interaction. It also has a visual display as the evaluation of the first iteration found that some visual feedback would be beneficial for users. However, the interface can be used without any visual interaction.

The amblr is also interpreted as an art installation which was visited by approximately 1000 people over 5 days. The installation mapped the virtual space created by the amblr to a physical space. To control movement within the space, the user only had to walk around a room. The instant usability of this interface allowed people to explore the music without needing to first learn how to use the interface. The enjoyment and ease of use of the art installation showed that spatial auditory display for exploring a collection of music can be effective and engaging.

7.2 Further Research Questions

While this thesis answered many questions, there remains further work to be done. Additional research questions arose as work progressed. They are discussed in further detail below and are divided into further work applicable to virtual Ambisonics, acoustics modelling, and interfaces for music discovery.

Virtual Ambisonics Chapter 3 examined how the localisation accuracy and the reversal rates with virtual Ambisonics compares to measured HRTFs, but did not investigate whether virtual Ambisonics influences the externalisation of sounds sources. Brookes and Treble [2005] found that symmetrical HRTFs can decrease externalisation, so this should be examined further

While this thesis focused on minimal equipment and did not utilise head tracking, the effects of head tracking with virtual Ambisonics should also be studied. Virtual Ambisonics is well suited for head tracking as rotations can be performed in the B-format domain as matrix multiplications. If individualised HRTFs are available, then the study performed by Begault et al. [2001] can be repeated with virtual Ambisonics. They compared how individualised HRTFs, non-individualised HRTFs, head tracking, and reverberation affect localisation and externalisaion.

Convolution reverberation of B-format IRs is one means to create artificial reverberation, but it may not be the most appropriate approach for all applications. A more flexible approach which does not strive to accurately reproduce a specific space may be desired. There has been some preliminary work by Anderson and Costello [2009] which looks into how structures such as FDNs could implemented in the B-format domain.

Acoustics Modelling Additional evaluation of the different components of the full interactive room model is needed. The mixing time can be evaluated in a similar method to [Lindau et al., 2010]. The study determined the perceptual mixing time of nine rooms through listening tests. Participants listened to a drum sample convolved with a binaural IR. Head tracking was used, so participants could rotate their heads and the source would remain static through cross-fading between multiple binaural IRs. The tail of the IRs was kept static and only the early portion was updated. Participants adjusted how much of the binaural IR was kept static and the mixing time was then the point in time that the participants could not hear any artefacts or clicks when rotating their head. To evaluate the kurtosis metric described here, the data set in Appendix B can be used. Head rotations from head tracking could easily be incorporated with virtual Ambisonics instead of requiring multiple measured binaural IRs as done by Lindau et al. [2010].

The effects of interpolating early reflections needs to also be evaluated with listening tests to determine the perceptual effects of the errors incurred. The primary remaining question is: what is the maximum distance between a pair of measured IRs that an IR can be interpolated without perceptible effects? It is expected that the answer will be dependent on the distance between the source and receiver. The just-noticeable error of the nearest source would then dictate the spatial sampling of a room.

As the intent of the analysis-synthesis model is to create reverberation that is indistinguishable from the measured IRs, listening tests are needed to perceptually verify the results. The question of note is whether the error introduced by the model can be perceived by listeners. This can be verified by a standard listening test traditionally used to test audio codecs and hardware, the ABX test. Listeners are presented with three audio files convolved with the same source material: a measured IR (file A), the modeled IR at the same position as the measured (file B), and a third audio file which is randomly either the measured or modelled IR (file X). The listener needs to determine whether the third audio file is the measured or modelled IR. When the percentage of correct answers is no better than the expected number when guessing, it can be implied that there is no perceptible difference between the measured and modelled IRs. That is, when the correct answers make up only 50% of the results it can be assumed that the listener cannot tell the difference between the two audio files and is only guessing.

The next step to developing the room model described in Chapter 4 is to build the full interactive system. A few implementation questions remain, primarily how to best cross-fade between early reflections, but much research has been done to address this within binaural research which cross-fading between HRTFs.

To evaluate the full interactive system:

1. Choose a selection of locations and orientations within a room.

2. Randomly select a location and orientation for each trial

3. Have the participant try to approach the sound source as quickly as possible.

4. Observe and time each trial. See if users consistently have the same issues such as difficulties when starting from a particular location or orientation. Compare this to areas where the model performs better or worse.

5. If possible, repeat this exercise in the same physical room that is modeled. With reasonable health and safety, blindfold the participant have them try to locate the source.

6. Evaluate whether the participant's actions differ when locating the source in the physical or modelled room.

Even if the evaluation cannot be repeated in the physical room, participants navigating towards a virtual sound source show that directional acoustic cues have been modelled.

Music Discovery The amblr's initial evaluation determined its usability for navigating a music collection, but the second iteration has not yet been evaluated. An evaluation should compare the effectiveness of the amblr to a commonly used interface that does not incorporate spatial auditory display.

The evaluation should entail the following steps:

1. Choose a series of tasks that require a participant to select a song. Examples include: listening to a song and being asked to find that song in a collection and choosing a song that would be ideal for a certain situation such as the gym. The tasks should be a mixture of those that have a definite correct or incorrect answer, and ones with a more subjective outcome where the participant judges how well they think they performed.

2. Populate a standard music collection interface such as iTunes with the same collection of music.

3. Ask the participant to perform the same tasks on each interface.

4. Randomise the order of the tasks and interfaces.

The variety of tasks lets two factors be evaluated: the time it takes to perform a search and how satisfied a user is with the result of a browsing task. Searching and browsing are different tasks and should be examined separately. By performing both types of tasks on both systems the evaluation compares how the amblr performs versus the standard interface.

The amblr implements what Cohen [1991] calls the museum metaphor while the art installation uses a cocktail party metaphor. Spatial auditory displays for music discovery most often use the museum metaphor, but the cocktail party metaphor could be an effective way to incorporate more social interaction into music discovery. The virtual space mapped onto the physical space created by the art installation could be shared by multiple users. Individual users could be represented with audio by having a song mapped to their movements. Other users would then hear that song in the virtual space. The physical space could also be removed so that users only move in a shared virtual space.

The third metaphor described by Cohen [1991], the theatre metaphor, may be a better method for presenting audio within the amblr when used as a music collection navigation application. In the theatre metaphor, the user does not move, but the audio content moves and is presented from different locations. The user is then seated at a personalised show. Evaluation of the amblr indicates that the museum metaphor is best with visual feedback, and the most effective auditory displays discussed in Chapter 5 that use the museum perspective also use visual displays. It is difficult to convincingly create the illusion of self-movement with audio only, so an approach without self-movement as done by Walker and Brewster [1999] may be best.

A third design iteration for the amblr should remove self-movement and investigate using auditory display directly as a means to navigate a search query. If pair with a audio content search engine and used to find similar audio files within a large collection, the amblr could have direct commercial applications within professional audio production studios and home entertainment.

The physical interface of the art installation of the amblr was appealing to the general public and further tangible and physical interfaces for music discovery should be explored. The physical aspect of the interface gave transparency to the spatialisation of the audio and also a unique mode of interaction. Augmented reality is becoming increasingly popular. It is beginning to be integrated into more consumer products showing that people like tangible interactions with real objects and completing tasks outside of a solely virtual space. However, augmented reality almost solely focuses on incorporating computer graphics and vision into an interface and seldom considers audio. Translating the art installation from the gallery to the street could be a transformative innovation.

Spatial auditory display can create more immersive and enjoyable applications. This thesis has shown that it is possible to create virtual environments that could be used for a variety of tasks and without dedicated hardware. Further work will more fully develop the applications investigated here and exploit the full potential of spatial auditory display.

References

The Human-Computer Interaction Handbook: fundamentals, evolving technologies and emerging applications, chapter Non-Speech Auditory Output, pages 220–239. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, 2001.

J. S. Abel and P. Huang. A simple, robust measure of reverberation echo density. In *Proc. of the 121st AES Conv.*, San Francisco, October 2006.

J. S. Abel, J. Rick, P. Huang, M. Kolar, J. O. Smith III, and J. M. Chowning. On the acoustics of the underground galleries of ancient Chavín de huántar, Peru. In *Proc. of Acoustics 08*, 2008.

P. D. Adamczyk. Seeing sounds: exploring musical social networks. In *MULTIMEDIA '04: Proc. of the 12th Annual ACM Int. Conf. on Multimedia*, pages 512–515, New York, NY, 2004. doi: 10.1145/1027527.1027651.

T. Ajdler. *The Plenacoustic function and its applications.* PhD thesis, École Polytechnique Fédérale De Lausanne, Lausanne, Switzerland, October 2006.

T. Ajdler, C. Faller, L. Sbaiz, and M. Vetterli. Interpolation of head related transfer functions considering acoustics. In *Proc. of 118th AES Conv.*, Barcelona, Spain, May 2005.

V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *Proc. of IEEE WASPAA: Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

S. Ali and P. Aarabi. A cyclic interface for the presentation of multiple music files. *IEEE Trans. on Multimedia*, 10(5):780–793, August 2008.

J. Anderson and S. Costello. Adapting artificial reverberation architectures for B-format signal processing. In *Proc. of the Ambisonics Symposium 2009*, Graz, Austria, June 2009.

C. Ardito, P. Buono, M. Costbile, R. Lanzilotti, T. Pederson, and A. Piccinno. Experiencing the past through the senses: an M-learning game at archaeological parks. *IEEE Multimedia*, 15(4):76–81, Oct.-Dec. 2008.

B. Arons. A review of the cocktail party effect. *J. of the American Voice I/O Society*, 12:35–50, 1992.

B. Arons. SpeechSkimmer: interactively skimming recorded speech. In *UIST '93: Proc. of the 6th Annual ACM Symposium on User Interface Software and Technology*, pages 187–196, 1993. doi: 10.1145/168642.168661.

B. Arons. Speechskimmer: a system for interactively skimming recorded speech. *ACM Trans. on Computer-Human Interaction*, 4(1):3–38, 1997. doi: 10.1145/244754.244758.

B. Arons, C. Binding, K. Lantz, and C. Schmandt. The VOX audio server. In *2nd IEEE Comsoc Int. Multimedia Communications Workshop*, Ottawa, Ontario, April 1989.

J.-J. Aucouturier and M. Sandler. Finding repeating patterns in acoustic musical signals: applications for audio thumbnailing. In *Proc. of AES 22nd Int. Conf.on Virtual, Synthetic, and Entertainment Audio*, June 2002.

S. Barrass and C. Frauenberger. A communal map of design in auditory display. In *ICAD '09*, 2009.

L. Barrington, R. Oda, and G. Lanckriet. Smarter than Genius: human evaluation of music recommender systems. In *Proc. of ISMIR'09: 10th Int.Society for Music Information Retrieval Conf.*, pages 357–362, Kobe, Japan, October 2009.

D. R. Begault. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *J. of the Audio Engineering Society*, 40(11):895–904, November 1992.

D. R. Begault. *3-D sound for virtual reality and multimedia.* NASA, 2000.

D. R. Begault, E. M. Wenzel, and M. R. Anderson. Direct comparison or the impact of head tracking, reverberation, and individualised head-related transfer functions on the spatial perception of a virtual speech source. *J. of the Audio Engineering Society*, 49 (10):904–916, October 2001.

R. Ben-Hador and I. Neoran. Capturing manipulation and reproduction of sampled acoustic impulse responses. In *Proc. the 117th AES Conv.*, 2004.

L. L. Beranek. Concert hall acoustics-1992. *J. of the Acoustical Society of America*, 92 (1):1–39, July 1992.

L. L. Beranek. *Concert halls and opera houses.* Springer-Verlag, 2nd edition, 2004.

A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *J. of the Acoustical Society of America*, 93(5):2764–2778, 1993. doi: 10.1121/1.405852.

S. Bertet, J. Daniel, L. Gros, E. Parizet, and O. Warusfel. Investigation of the perceived spatial resolution of higher order ambisonics sound fields: a subjective evaluation involving virtual and real 3D microphones. In *Proc. of AES 30th Int. Conf. on Intellient Audio Environments*, 2007.

M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg. Earcons and icons: their structure and common design principles. *Human-Computer Interaction*, 4(1):11–44, 1989.

J. Blauert. *Spatial hearing: the psychophysics of human sound localization.* MIT Press, 3rd edition, 1997.

B. Blesser. An interdisciplinary synthesis of reverberation viewpoints. *J. of the Audio Engineering Society*, 49(10):867–903, October 2001.

B. Blesser and L.-R. Salter. *Spaces Speak: Are you listening?* MIT Press, 2007.

C. Borß and R. Martin. Interactive auditory virtual environments for mobile devices. In *MobileHCI '08: Proc.of the 10th Int. Conf. on Human Computer Interaction with Mobile Devices and Services*, pages 487–488, New York, NY, USA, 2008. ACM. doi: 10.1145/1409240.1409322.

J. Bradley. Auditorium acoustics measurements from pistol shots. *J. of the Acoustical Soc. of America*, 80(1):199–205, July 1986.

J. Bradley. Comparison of concert hall measurements of spatial impression. *J. of the Acoustical Society of America*, 96(6):3525–3535, December 1994.

E. Brazil and M. Fernström. Audio information browsing with the Sonic Browser. In *CMV '03: Proceedings of IEEE Conf. on Coordinated and Multiple Views in Exploratory Visualization*, 2003.

E. Brazil, M. Fernström, G. Tzanetakis, and P. R. Cook. Enhancing sonic browsing using audio information retrieval. In *Proc. of ICAD '02: Int. Conf. on Auditory Display*, Kyoto, Japan, July 2002.

E. Brazil, M. Fernström, and J. Bowers. Exploring concurrent auditory icon recognition. In *Proc. of ICAD '09: Int. Conf. on Auditory Display*, Copenhagen, Denmark, May 2009.

A. S. Bregman. *Auditory Scene Analysis: the perceptual organization of sound.* MIT Press, Cambridge, MA, 2nd edition, 1999.

T. Brookes and C. Treble. The effect of non-symmetrical left/right recoring pinnae on the perceived externalisation of binaural recordings. In *Proc. of 118th AES Conv.*, Barcelona, Spain, May 2005.

S. Browne. Hybrid reverberation algorithm using truncated impulse response convolution and recursive filtering. Master's thesis, University of Miami, Coral Gables, FL, 2001.

O. Celma and P. Cano. From hits to niches? or how popular artists can bias music recommendation and discovery. In *Proc. of 2nd Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition (ACM KDD)*, Las Vegas, Nevada, USA, August 2008.

C. I. Cheng and G. H. Wakefield. Introduction to head-related transfer functions (HRTFs): representations of HRTFs in time, frequency, and space. *J. of the Audio Engineering Society*, 49(4):231–249, April 2001.

E. C. Cherry. Some experiments on the recognition of speech, with one and two ears. *JASA*, 25(5):979–979, September 1953.

R. Clifton. Breakdown of echo suppresion in the precedence effect. *J. of the Acoustical Society of America*, 82:1834–1835, 1987.

M. Cohen. Multidimensional audio window management. *Int. J. Man-Machine Studies*, 34:319–336, 1991.

G. Coleman. Mused: navigating the personal sample library. In *Proc. of ICMC: Int. Computer Music Conf.*, Copenhagen, Denmark, August 2007.

S. J. Cunningham, N. Reeves, and M. Britland. An ethnographic study of music information seeking: implications for the design of a music digital library. In *JCDL '03: Proc. of the 3rd ACM/IEEE-CS Joint Conf. on Digital Libraries*, pages 5–16, Houston, Texas, USA, 2003.

B. Dalenbäck and M. Strömberg. Real time walkthrough auralization - the first year. In *Proc. - Institute of Acoustics*, volume 28, Pt. 2, 2006.

J. Daniel. Spatial sound encoding including near field effect: introducting distance coding filters and a viable, new ambisonic format. In *Proc. oc AES 23rd Int. Conf. on Signal Processing in Audio Recording and Production*, Copenhagen, May 2003.

J. Daniel, J.-B. Rault, and J.-D. Polack. Ambisonics encoding of other audio formats for multiple listening conditions. In *Proc. of 105th AES Conv.*, San Francisco, September 1998.

D. de Vries, E. Hulsebos, and J. Baan. Spatial fluctuations in measures for spaciousness. *J. of the Acoustical Society of America*, 110(2):947–954, August 2001.

A. Farina. Simultaneous measurement of impulse response and distortion with a swept sine technique. In *Proc. of the 108th AES Conv.*, Paris, France, February 2000.

A. Farina and R. Ayalon. Recording concert hall acoustics for posterity. In *Proc. of AES 24th Int. Conf. on Multichannel Audio*, 2003.

P. Fellgett. Ambisonics. Part one: general system description. *Studio Sound*, pages 20–22, 40, August 1975.

M. Fernström. Reflections on Sonic Browsing: comments on Fernström and McNamara, ICAD 1998. *ACM Trans.on Applied Perception*, 2(4):500–504, October 2005.

M. Fernström and E. Brazil. Sonic browsing: an auditory tool for multimedia asset management. In *Proc. of ICAD '01: Internation Conf. on Auditory Display*, pages 132–135, Espoo, Finland, August 2001.

M. Fernström and C. McNamara. After direct manipulation - direct sonification. In *Proc. of ICAD '98: Int. Conf. on Auditory Display*, 1998.

R. T. Fielding. *REST: Architectural Styles and the Design of Network-based Software Architectures.* Doctoral dissertation, University of California, Irvine, 2000.

J. Frank, T. Lidy, E. Peiszer, R. Genswaider, and A. Rauber. Ambient music experience in real and virtual worlds using audio similarity. In *AME '08: Proc.of the 1st ACM Int.Workshop on Semantic Ambient Media Experiences*, pages 9–16, Vancouver, British Columbia, Canada, 2008.

C. Frauenberger and T. Stockman. Auditory display design – and investigation of a design pattern approach. *Int. Journal Human-Computer Studies*, 67:907–922, 2009.

W. Gardner. *Applications of Digital Signal Processing to Audio and Acoustics*, chapter "Reverberation algorithms", pages 85–131. Kluwer Academic Pub., MA, 1998.

W. G. Gardner and K. Martain. HRTF measurements of a KEMAR dummy-head microphone. Technical report, MIT Media Lab, May 1994.

W. W. Gaver. Auditory icons: using sound in computer interfaces. *Human-Computer Interaction*, 2(2):167–177, 1986. doi: 10.1207/s15327051hci0202_3.

W. W. Gaver. The sonicfinder: an interface that uses auditory icons. *Human-Computer Interaction*, 4(1):67–94, 1989.

M. A. Gerzon. Periphony: with-height sound reproduction. *J. of the Audio Engineering Society*, 21(1):3–10, January/February 1973.

M. A. Gerzon. The design of precisely conincident microphone arrays for stereo and surround sound. In *Proc. of 50th AES Conv.*, London, UK, March 1975.

M. A. Gerzon. Unitary (enery preserving) multichannel networks with feedback. *Electronics Letters*, 12(11):278–279, 1976.

M. A. Gerzon. General metatheory of auditory localisation. In *Proc. of 92nd AES Conv.*, Vienna, Austria, March 1992.

J. D. Goldenson. Beat browser. Master's thesis, MIT, 2007.

L. S. Goodfriend and J. H. Beaumont. The development and application of synthetic reverberation systems. *J. of the Audio Engineering Society*, 7(4):228–250, October 1959.

A. B. Greenblatt, J. S. Abel, and D. P. Berners. A hybrid reverberation crossfading technique. In *Proc. of IEEE Conf. on Acoustics, Speech and Signal Processing*, pages 429–432, Dallas, TX, March 2010.

H. Haas. The influence of a single echo on the audibility of speech. *J. of the Audio Engineering Society*, 20(2):146–159, May 1972.

M. Hamanaka and S. Lee. Music Scope Headphones: natural user interface for selection of music. In *Proc. of ISMIR'06: 7th Int. Society for Music Information Retrieval Conf.*, 2006.

W. M. Hartmann. Localization of sound in rooms. *J. of the Acoustical Society of America*, 74(5):1380–1391, November 1983.

K. Hartung, J. Braasch, and S. J. Sterbing. Comparison of difference methods for the interpolation of head-related transfer functions. In *Proc. of the 16th AES Int. Conf.*, 1999.

S. Heise, M. Hlatky, and J. Loviscach. SoundTorch: Quick browsing in large audio collections. In *Proc. of AES 125th Conv.*, San Francisco, CA, October 2008.

S. Heise, M. Hlatky, and J. Loviscach. Aurally and visually enhanced audio search with SoundTorch. In *CHI '09: Proc. of the 27th int. conf.e extended abstracts on Human factors in computing systems*, pages 3241–3246, Boston, MA, USA, April 2009. doi: 10.1145/1520340.1520465.

A. J. Heller, R. Lee, and E. M. Benjamin. Is my decoder ambisonic In *Proc. of 125th AES Conv.*, San Francisco, CA, October 2008.

T. Hermann and A. Hunt. Guest Editors' Introduction: an introduction to interactive sonification. *IEEE Multimedia*, 12:20–24, 2005. ISSN 1070-986X. doi: 10.1109/MMUL.2005.26.

W. Heuten, N. Henze, and S. Boll. Interactive exploration of city maps with auditory torches. In *Proc. of CHI '07: CHI '07 Extended Abstracts on Human Factors in Computing Systems*, San Jose, CA, USA, April 2007.

T. Hidaka, Y. Yamada, and T. Nakagawa. A new definition of boundary point between early reflections and late reverberation in room impulse responses. *J. of the Acoustical Society of America*, 122(1):326–332, July 2007.

J. Hiipakka and G. Lorho. A spatial audio user interface for generating music playlists. In *Proc. of ICAD '03: Int. Conf. on Auditory Display*, Boston, MA, USA, July 2003.

M. Holters, T. Corbach, and U. Zölzer. Impulse response measurement techniques and their applicability in the real world. In *Proc. of DAFx '09: 12th Int. Conf. on Digital Audio Effects*, Como, Italy, September 2009.

D. M. Howard and J. Angus. *Acoustics and Psychoacoustics.* Focal Press, 3rd edition, 2006.

D. M. Huber and R. E. Runstein. *Modern Recording Techniques.* Focal Press, 5th edition, 2001.

J. Huopaniemi, L. Savioja, and T. Takala. DIVA virtual audio reality system. In *Proc. ICAD'96: Int. Conf. Auditory Display*, pages 111–116, 1996.

C. Huszty, Németh, P. Baranyi, and R. Augustinovicz. Measurement-based fuzzy interpolation of room impulse responses. In *Proc. Acoustics 08*, pages 5827–5832, Paris, France, June 2008.

ISO 3382. Acoustics-measurements of the reverberation time of rooms with reference to other acoustical parameters, 1997.

R. Jeffries and H. Desurvire. Usability testing vs. heuristic evaluation: was there a contest? *SIGCHI Bull.*, 24(4):39–41, 1992. doi: 10.1145/142167.142179.

M. Jeub, M. Schäfer, and P. Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proc. of IEEE 16th Int. Conf. on Digital Signal Processing, 2009*, July 2009.

J.-M. Jot. Digital delay networks for designing artificial reverberators. In *Proc. of 90th AES Conv.*, Paris, France, February 1991.

J.-M. Jot, L. Cerveau, and O. Warusfel. Analysis and synthesis of room reverberation based on a statistical time-frequency model. In *Proc. of 103rd AES Conv.*, New York, NY, September 1997.

E. Jovanov, K. Wegner, V. Radivojević, M. S. Quinn, and D. B. Karron. Tactical audio and acoustic rendering in biomedical applications. *IEEE Trans. on Information Technology in Biomedicine*, 3(2):109–118, June 1999.

M. Kleiner, B.-I. Dalenback, and P. Svensson. Auralization-an overview. *J. of the Audio Engineering Society*, 41(11):861–875, November 1993.

P. Knees, M. Schedi, T. Pohle, and G. Widmer. An innovative three-dimensional user interfacefor exploring music collections enriched with meta-information from the web. In *MULTIMEDIA '06: Proc. of the 14th annual ACM int.l conf. on Multimedia*, pages 17–24, Santa Barbara, CA, USA, 2006. doi: 10.1145/1180639.1180652.

M. Kobayashi and C. Schmandt. Dynamic Soundscape: mapping time to spae for audio browsing. In *CHI '97: Proc. of the SIGCHI conf. on human factors in computing systems*, pages 194–201, March 1997. doi: 10.1145/258549.258702.

G. Kramer. *Auditory Display: Sonification, audification, and auditory interfaces.* Addison-Wesley, 1994.

H. Kuttruff. *Room acoustics.* Spon Press, London, 4th edition, 2000.

M.-V. Laitinen. Binaural reproduction for directional audio coding. Master's thesis, Helsinki University of Technology, Helsinki, Finland, 2008.

M. Levy and M. Sandler. A semantic spacce for music derived from social tags. In *Proc. of ISMIR'07: 8th Int. Society for Music Information Retrieval Conf.*, pages 411–416, 2007.

A. Lindau, L. Kosanke, and S. Weinzierl. Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses. In *Proc. of 128th AES Conv.*, London, UK, May 2010.

T. Lokki and L. Savioja. Evaluation of auralization results. In *Forum Acusticum 2005*, Budapest, Hungary, 2005.

A. Lomax. *Cantometrics: An Approach To The Anthropology Of Music.* The University of California, 1976. accompanied by 7 cassettes.

G. Lorho, J. Marila, and J. Hiipakka. Feasibility of multiple non-speech sounds presentation using headphones. In *Proc. of ICAD '01: Int. Conf. on Auditory Display*, Espoo, Finland, August 2001.

D. Lübbers. Sonixplorer: Combining visualization and auralization for content-based exploration of music collections. In *Proc. of ISMIR'05: 6th Int. Society for Music Information Retrieval Conf.*, pages 590–593, London, UK, 2005.

D. Lübbers and M. Jarke. Adaptive multimodal exploration of music collections. In *Proc. of ISMIR'09: 10th Int. Society for Music Information Retrieval Conf.*, pages 195–200, Kyoto, Japan, 2009.

D. Malham. Experience with large area 3-D ambisonics sound systems. In *Proc. of Institute of Acoustics Autumn Conf. on Reproduced Sound 8*, 1992.

D. Malham. Approaches to spatialisation. *Organised Sound*, 3(2):167–177, 1998.

D. Malham. Higher order ambisonics systems. Technical report, 2009. URL `http://www.york.ac.uk/inst/mustech/3d_audio/higher_order_ambisonics.pdf`.

D. Malham and A. Myatt. 3-D sound spatialization using Ambisonic techniques. *Computer Music J.*, 19(4):58–70, Winter 1995.

N. Mariette, B. F. G. Katz, K. Bousetta, and O. Guillerminet. Sounddelta: a study of audio augments reality using WiFi-distributed Ambisonic cell rendering. In *Proc. of 128th AES Conv.*, London, UK, May 2010.

C. Masterson, G. Kearney, and F. Boland. Acoustic impulse response interpolation for multichannel systems using dynamic time warping. In *Proc. of AES 35th Int. Conf. on Audio for Games*, London, UK, February 2009.

D. McGrath and A. Reilly. Creation, manipulation and playback of soundfields with the Huron digital audio convolution workstation. In *Int. Symposium on Signal Processing and its Applications*, Gold Coast, Australia, August 1996.

A. McKeag and D. McGrath. Sound field format to binaural decoder with head tracking. In *AES 6th Australian Regional Conv.*, Melbourne, Australia, September 1996.

H. Moller, M. F. Sorensen, C. B. Jensen, and D. Hammershoi. Binaural technique: do we need individual recordings *J. of the Audio Engineering Society*, 44(6):451–469, June 1996.

B. C. Moore. *An Introduction to the Psychology of Hearing.* Academin Press, 5th edition, 2003.

J. A. Moorer. About this reverberation business. In *Computer Music J.*, volume 2, Summer, 1979.

M. Morimoto, M. Jinya, and K. Nakagawa. Effects of frequency characteristics of reverberation time on listener envelopment. *J. of the Acoustical Society of America*, 122(3):1611–1615, September 2007.

C. E. R. A. Moura and S. L. Campos. Some notes on artificial reverberation. *Journal of the Audio Engineering Society*, 5(4):182–186, October 1959.

D. Murphy, M. Beeson, S. Shelley, and A. Moore. Hybrid room impulse response synthesis in digital waveguide mesh based room acoustics simulation. In *Proc. of DAFx '08: 11th Int. Conf. on Digital Audio Effects*, Helsinki, Finland, September 2008.

D. T. Murphy. Archaeological acoustic space measurement for convolution reverberation and auralization applications. In *Proc. of DAFx '06: 9th Int. Conf. on Digital Audio Effects*, September 2006.

C. L. Nikias and A. P. Petropulu. *Higher-order spectra analysis: a nonlinear signal processing framework.* Prentice Hall, 1993.

T. Nishino, S. Kajita, K. Takeda, and F. Itakura. Interpolating head related transfer functions in the median plane. In *Proc. of IEEE WASPAA: Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 167–170, 1999.

M. Noisternig, T. Musil, A. Sontacchi, and R. Holdrich. 3D binaural sound reproductions using a virtual ambisonic approach. In *Proc. of VECIMS: Int. Symposium on Virtual Environments, Human-Computer Interfaces, and Measurement Systems*, Lugano, Switzerland, July 2003a.

M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich. A 3D ambisonic based binaural sound reproduction system. In *AES 24th Int. Conf. on Multichannel Audio*, 2003b.

E. Pampalk. Islands of music: analysis, organization, and visualization of music archives. Master's thesis, Vienna University of Technology, 2001.

S. Pauws, D. Bouwhuis, and B. Eggen. Programming and enjoying music with your eyes closed. In *CHI '00: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 376–383. ACM, 2000. doi: 10.1145/332040.332460.

J.-D. Polack. Playing billiards in the concert hall: the mathematical foundations of geometrical room acoustics. *Applied Acoustics*, 38:235–244, 1993.

V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *J. of the Audio Engineering Society*, 45(6):456–466, June 1997.

V. Pulkki. *Spatial sound generation and perception by amplitude panning techniques.* PhD thesis, Helsinki University of Technology, Espoo, Finland, 2001a.

V. Pulkki. Evaluating spatial sound with binaural auditory model. In *Proc. of ICMC: Int. Computer Music Conf.*, 2001b.

D. Radford. Hybrid artificial reverberation algorithms for VST. Master's thesis, University of York, York, UK, 2003.

L. Rayleigh. On our perception of sound direction. *Phil. Mag.*, 13(6th series):214–232, 1907.

A. Reilly and D. McGrath. Using auralisation for creating animated 3-D sound fields across multiple speakers. In *Proc. of 99th AES Conv.*, New York, NY, October 1995a. 99th AES Conv.

A. Reilly and D. McGrath. Real-time auralization wih head tracking. In *Proc. of AES 5th Australian Regional Conv.*, Sydney, Austrailia, April 1995b.

P. Rubak. Evaluation of artificial reverberation decay quality. In *118th AES Conv.*, Barcelona, Spain, May 2005.

F. Rumsey. *Spatial Audio.* Focal Press, 2001.

W. Sabine. *Collected Papers on Acoustics: Wallace Clement Sabine.* Peninsula Publishing, Los Altos, CA, 1992.

C. Schmandt. Audio Hallway: a virtual acoustic environment for browsing. In *UIST '98: Proc. of the 11th Annual ACM Symposium on User Interface Software and Technology*, pages 163–170, San Francisco, California, United States, 1998. doi: 10.1145/288392.288597.

C. Schmandt and A. Mullins. AudioStreamer: exploiting simultaneity for listening. In *CHI '95: Conf. Companion on Human Factors in Computing Systems*, pages 218–219, Denver, Colorado, United States, May 1995. doi: 10.1145/223355.223533.

M. Schroeder. Natural sounding artificial reverberation. *J. of the Audio Engineering Society*, 10(3):219–223, October 1962.

M. Schroeder. New method of measuring reverberation time. *J. of the Acoustical Society of America*, 37(3):409–412, March 1965.

M. R. Schroeder. The "Schroeder freqency" revisited. *J. of the Acoustical Society of America*, 99(5):3240–3241, May 1996.

F. S. Scott and A. Roginska. Room-dependent preference of virtual surround sound. In *Proc. of 124th AES Conv.*, Amsterdam, The Netherlands, May 2008.

B. U. Seeber and H. Fastl. Subjective selection of non-individual head-related transfer functions. In *Proc. of ICAD '03: Int. Conf. on Auditory DisplayICAD '03*, Boston, MA, USA, July 2003.

C. W. Sheeline. *An investigation of the effects of direct and reverberant signal interactions on auditory distance perception.* PhD thesis, CCRMA, Stanford, Stanford, California, USA, 1982.

B. Shinn-Cunningham. Applications of virtual auditory displays. In *Proceedings of the 20th Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, pages 1105–1108, 1998.

B. Shinn-Cunningham. Learning reverberation: considerations for spatial audio displays. In *Proc. of ICAD '00: Int. Conf. on Auditory Display*, 2000.

B. Shneiderman. *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Addison-Wesley, 3rd edition, 1997.

A. Southern, J. Wells, and D. Murphy. Rendering walk-through auralisations using wave-based acoustical models. In *Proc. of 17th European Sigan Processing Conf. (EUSIPCO 2009)*, Glasgow, Scotland, August 2009.

R. Stewart and D. Murphy. A hybrid artificial reverberation algorithm. In *122nd AES Conv.*, Vienna, Austria, May 2007.

C. Travis. A virtual perspective on headphone audio. In *UK 11th Conf.: Audio for New Media*, March 1996.

G. Tzanetakis and P. R. Cook. MARSYAS: a framework for audio analysis. *Organised Sound*, 4(3):169–175, December 2000.

G. Tzanetakis and P. R. Cook. MARSYAS3D: a prototype audio browser-editor using a large scale immersive visual and audio display. In *Proc. of ICAD '01: Internation Conf. on Auditory Display*, Espoo, Finland, August 2001.

J. Vanderkooy. Aspects of MLS measuring systems. *J. of the Audio Engineering Society*, 42(4):219–231, April 1994.

M. Vorländer. *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality.* Springer, 2008.

A. Walker and S. Brewster. Extending the auditory display space in handheld computing devices. In *Proc. of the Second Workshop on Human Computer Interaction with Mobile Devices*, 1999.

J. Y. C. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor. Evaluation of speech dereverberation algorithms using the MARDY database. In *Proc. of Int. Workshop on Acoustic Echo and Noise Control*, Paris, France, September 2006.

E. M. Wenzel. *Multimedia interface design*, chapter Three-dimensional virtual acoustic displays, pages 257–288. ACM, New York, NY, USA, 1992.

E. M. Wenzel. Research in virtual acoustic displays at NASA. In *SimTecT 96: the Simulation Technology and Training Conf.*, Melbourne, Australia, March 1996.

E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *JASA*, 94(1):111–123, July 1993.

G. Wersényi. Effect of emulated head-tracking for reducing localization errors in virutal audio simulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17 (2):247–252, 2009.

B. Wiggins. *An investigation into the real-time manipulation and control of three-dimensional sound fields.* PhD thesis, University of Derby, Derby, UK, 2004.

B. Wiggins, I. Paterson-Stephens, and P. Schillebeeckx. The analysis of multi-channel sound reproduction algorithms using HRTF data. In *19th Int. AES Surround Sound Conv.*, pages 111–123, Germany, 2001.

E. Zwicker and H. Fastl. *Psychoacoustics: facts and models.* Springer, 2nd edition, 1999.

Appendix A

Virtual Ambisonics Listening Test

This appendix documents the listening tests conducted in Section 3.4. Figure A.1 shows the graphical user interface used by participants to indicate the perceived position of the sound source. Figures A.2 and A.3 are the instructions that each participant received.

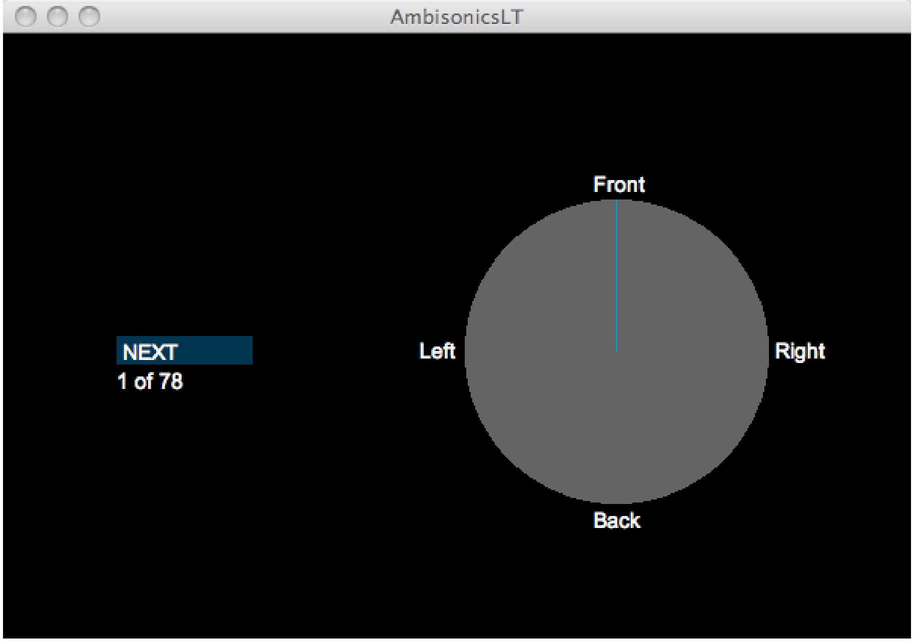
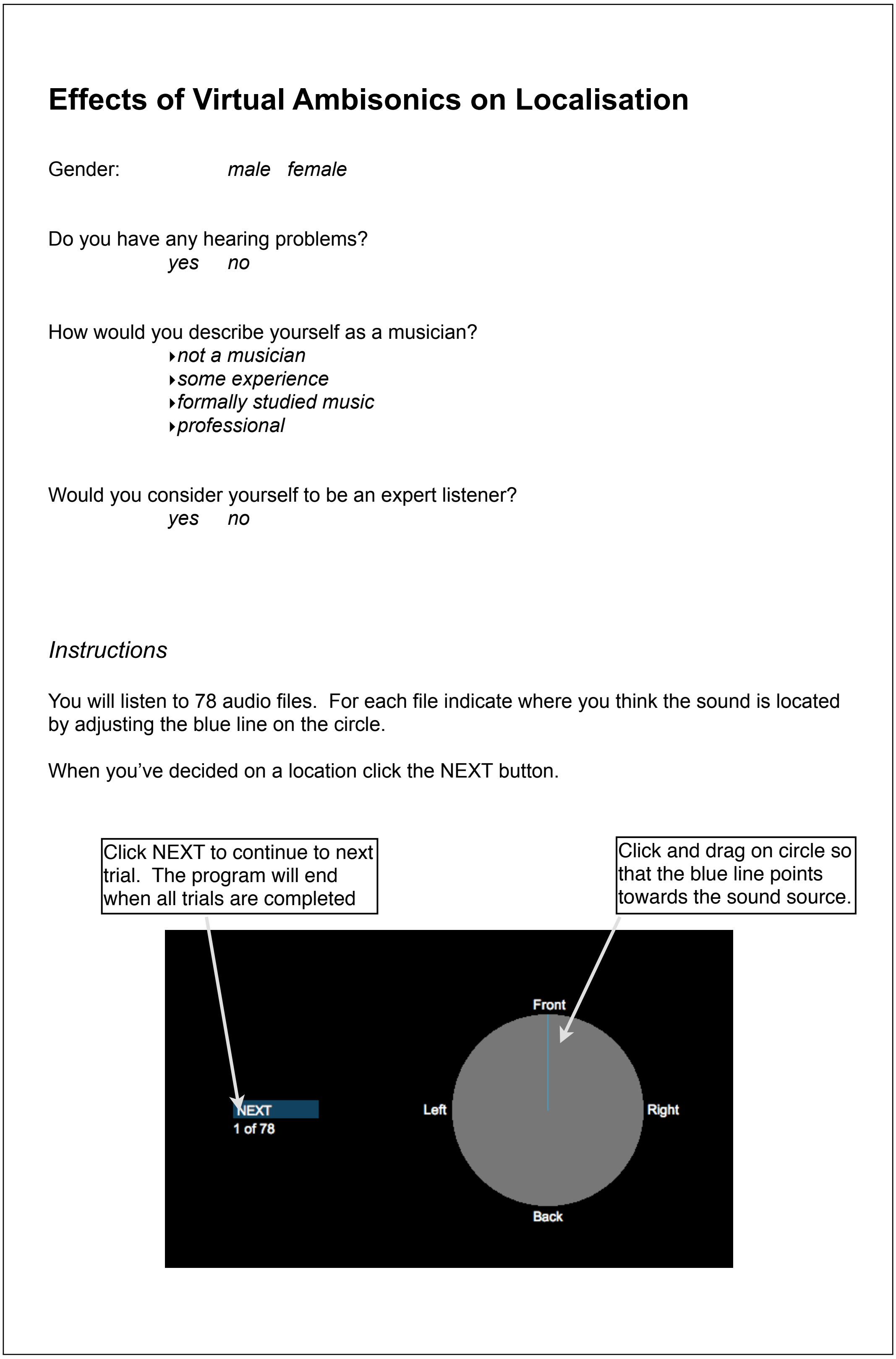


Figure A.1: The user interface for the listening tests in the evaluation of virtual Ambisonics.

The following tables list the raw data gathered from the listening tests including the perceived sound source azimuth, the reversal-corrected azimuth if appropriate, and the error.



Effects of Virtual Ambisonics on Localisation

Gender: *male female*

Do you have any hearing problems?
yes no

How would you describe yourself as a musician?
- *not a musician*
- *some experience*
- *formally studied music*
- *professional*

Would you consider yourself to be an expert listener?
yes no

Instructions

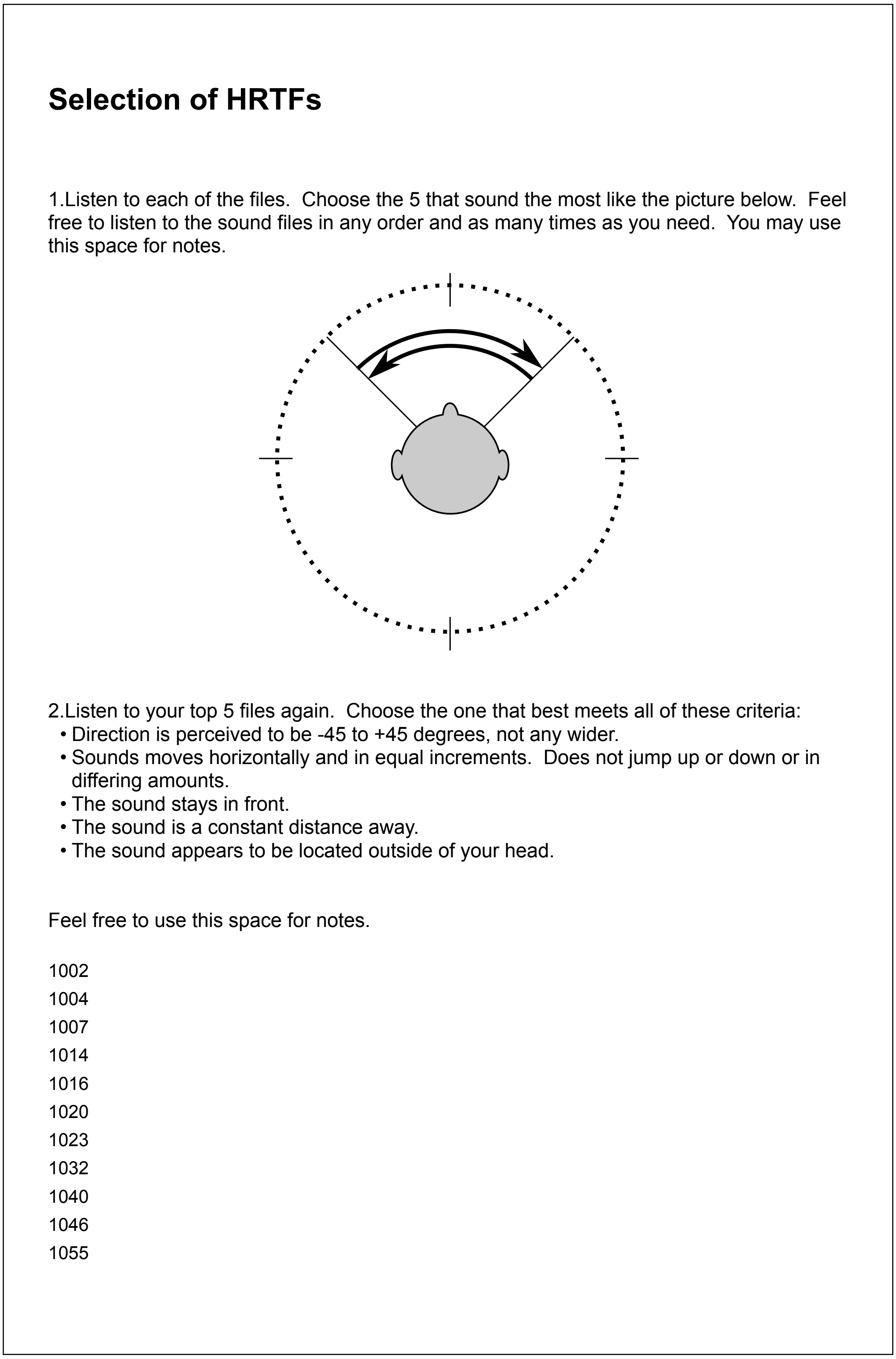
You will listen to 78 audio files. For each file indicate where you think the sound is located by adjusting the blue line on the circle.

When you've decided on a location click the NEXT button.

Click NEXT to continue to next trial. The program will end when all trials are completed

Click and drag on circle so that the blue line points towards the sound source.

Figure A.2: The first page of the instructions given to the listening test participants.



Selection of HRTFs

1.Listen to each of the files. Choose the 5 that sound the most like the picture below. Feel free to listen to the sound files in any order and as many times as you need. You may use this space for notes.

2.Listen to your top 5 files again. Choose the one that best meets all of these criteria:

- Direction is perceived to be -45 to +45 degrees, not any wider.
- Sounds moves horizontally and in equal increments. Does not jump up or down or in differing amounts.
- The sound stays in front.
- The sound is a constant distance away.
- The sound appears to be located outside of your head.

Feel free to use this space for notes.

1002
1004
1007
1014
1016
1020
1023
1032
1040
1046
1055

Figure A.3: The second page of the instructions given to the listening test participants.

Participant 0

Azimuth	*Direct*	*Rev*	*Error*	1st *On*	*Rev*	*Error*	1st *Off*	*Rev*	*Error*
0	140	40	40	151	29	29	147	33	33
30	131	49	19	135	45	15	160	20	10
45	110	70	25	117	63	18	105	75	30
60	103	77	17	119	61	1	107	73	13
90	89		1	129		39	97		7
120	92		28	118		2	128		8
150	126		24	120		30	103		47
180	168		12	179		1	150		30
225	246		21	243		18	211		14
270	258		12	205		65	208		62
300	271		29	252	288	12	211	329	29
330	204	336	6	192	348	18	185	365	35
345	182	358	13	155	45	60	134	46	61

Azimuth	2nd *On*	*Rev*	*Error*	2nd *Off*	*Rev*	*Error*	3rd *On*	*Rev*	*Error*
0	163	17	17	179	1	1	168	12	12
30	123	57	27	117	63	33	82		52
45	105	75	30	109	71	26	97	83	38
60	117	63	3	91	89	29	107	73	13
90	87		3	90		0	97		7
120	108		12	90		30	109		11
150	134		16	103		47	108		42
180	164		16	123		57	180		0
225	225		0	242		17	212		13
270	224		46	236		34	241		29
300	262	278	22	247	293	7	251	289	11
330	229	311	19	182	358	28	192	348	18
345	182	358	13	201	339	6	197	343	2

Participant 1

Azimuth	*Direct*	*Rev*	*Error*	1st *On*	*Rev*	*Error*	1st *Off*	*Rev*	*Error*
0	177	3	3	0		0	0		0
30	124	56	26	37		7	79		49
45	79		34	55		10	55		10
60	130	50	10	151	29	31	53		7
90	93		3	70		20	52		38
120	88	92	28	93		27	89	91	29
150	146		4	132		18	125		25
180	181		1	182		2	0	180	0
225	240		15	244		19	323	217	8
270	284		14	287		17	299		29
300	251	289	11	295		5	312		12
330	315		15	329		1	334		4
345	249	291	54	177	3	18	306		39

Azimuth	2nd *On*	*Rev*	*Error*	2nd *Off*	*Rev*	*Error*	3rd *On*	*Rev*	*Error*
0	336		24	178	2	2	0		0
30	9		21	47		17	148	32	2
45	108	72	27	100	80	35	74		29
60	101	79	19	114	66	6	95	85	25
90	72		18	90	83	7	93		3
120	64	116	4	92		28	91		29
150	56	124	26	123		27	139		11
180	0	180	0	178		2	0	180	0
225	233		8	247		22	230		5
270	299		29	253		17	240		30
300	264	276	24	240	300	0	226	314	14
330	306		24	314		16	308		22
345	202	338	7	190	350	5	327		18

Participant 2

Azimuth	*Direct*	*Rev*	*Error*	1^{st} *On*	*Rev*	*Error*	1^{st} *Off*	*Rev*	*Error*
0	179	1	1	11		11	182	-2	2
30	69		39	86		56	50		20
45	81		36	90		45	81		36
60	89		29	131	49	11	90		30
90	65		25	75		15	92		2
120	131		11	74	106	14	118		2
150	100		50	32	146	4	27	153	3
180	179		1	0	180	0	0	180	0
225	215		10	281	259	34	245		20
270	259		11	301		31	237		33
300	204	336	36	268	272	28	309		9
330	222	316	14	336		6	269	271	59
345	282		63	343		2	343		2

Azimuth	2^{nd} *On*	*Rev*	*Error*	2^{nd} *Off*	*Rev*	*Error*	3^{rd} *On*	*Rev*	*Error*
0	0		0	11		11	9		9
30	91	89	59	24		6	45		15
45	59		14	92	88	43	73		28
60	121	59	1	90		30	103		43
90	79		11	76		14	118		28
120	70	110	10	117		3	115		5
150	139		11	55	125	25	67	113	37
180	0	180	0	27	153	27	0	180	0
225	194		31	268		43	247		22
270	210		60	257		13	235		35
300	269	271	29	243	297	3	221	319	19
330	348		18	312		18	348		18
345	344		1	360		15	310		35

Participant 3

Azimuth	*Direct*	*Rev*	*Error*	1^{st} *On*	*Rev*	*Error*	1^{st} *Off*	*Rev*	*Error*
0	180	0	0	0		0	0		0
30	140	40	10	122	58	28	127	53	23
45	133	47	2	157	23	22	142	38	7
60	92	88	28	117	63	3	144	36	24
90	112		22	131		41	92		2
120	149		29	123		3	127		7
150	146		4	107		43	135		15
180	180		0	0	180	0	168		12
225	249		24	234		9	193		32
270	270		0	209		61	228		42
300	304		4	202	338	38	254	286	14
330	181	359	29	184	356	26	270		60
345	179	1	16	360		15	360		15

Azimuth	2^{nd} *On*	*Rev*	*Error*	2^{nd} *Off*	*Rev*	*Error*	3^{rd} *On*	*Rev*	*Error*
0	148	32	32	180	0	0	21		21
30	144	36	6	122	58	28	134	46	16
45	91	89	44	108	72	27	126	54	9
60	120	60	0	92	88	28	107	73	13
90	92		2	137		47	92		
120	116		4	123		3	150		30
150	140		10	149		1	157		7
180	180		0	172		8	0	180	0
225	255		30	270		45	360	180	45
270	217		53	235		35	220		50
300	284		16	270		30	238	302	2
330	325		5	182	358	28	360		30
345	181	359	14	360		15	215	325	20

Participant 4

Azimuth	*Direct*	*Rev*	*Error*	1st *On*	*Rev*	*Error*	1st *Off*	*Rev*	*Error*
0	187	-7	7	0		0	0		0
30	90		60	129	51	21	117	63	33
45	66		21	96	84	39	47		2
60	121	59	1	104	76	16	80		20
90	92		2	76		14	90		0
120	110		10	106		14	72	108	12
150	132		18	62	118	32	89	91	59
180	0	180	0	162		18	0	180	0
225	266		41	223		2	271	269	44
270	270		0	239		31	207		63
300	313		13	254	286	14	231	309	9
330	244	296	34	282		48	236	304	26
345	223	317	28	296		49	200	340	5

Azimuth	2nd *On*	*Rev*	*Error*	2nd *Off*	*Rev*	*Error*	3rd *On*	*Rev*	*Error*
0	0		0	0		0	185	-5	5
30	35		5	65		35	51		21
45	71		26	73		28	70		25
60	82		22	90		30	113	67	7
90	90		0	105		15	119		29
120	111		9	76	104	16	66	114	6
150	131		19	55	125	25	43	137	13
180	0	180	0	0	180	0	0	180	0
225	240		15	270		45	251		26
270	271		1	231		39	275		5
300	270		30	271		29	270		30
330	226	314	16	312		18	247	293	37
345	227	313	32	330		15	238	302	43

Participant 5

Azimuth	*Direct*	*Rev*	*Error*	1st *On*	*Rev*	*Error*	1st *Off*	*Rev*	*Error*
0	175	5	5	0		0	0		0
30	68		38	120	60	30	85		55
45	82		37	125	55	10	98	82	37
60	85		25	103	77	17	99	81	21
90	102		12	136		46	138		48
120	112		8	122		2	97		23
150	156		6	125		25	112		38
180	180		0	0	180	0	60	120	60
225	269		44	232		7	239		14
270	278		8	259		11	266		4
300	293		7	246	294	6	301		1
330	287		43	189	351	21	314		16
345	305		40	0		15	201	339	6

Azimuth	2nd *On*	*Rev*	*Error*	2nd *Off*	*Rev*	*Error*	3rd *On*	*Rev*	*Error*
0	145	35	35	113	67	67	0		0
30	100	80	50	77		47	82		52
45	116	64	19	90		45	71		26
60	90		30	99	81	21	100	80	20
90	91		1	99		9	105		15
120	99		21	107		13	141		21
150	104		46	113		37	111		39
180	180	113	67	171		9	356	184	4
225	245		20	239		14	216		9
270	285		15	235		35	260		10
300	289		11	248	292	8	252	288	12
330	312		18	284		46	307		23
345	360		15	186	354	9	300		45

Participant 6

Azimuth	*Direct*	*Rev*	*Error*	1st *On*	*Rev*	*Error*	1st *Off*	*Rev*	*Error*
0	18		18	341		19	23		23
30	77		47	28		2	12		18
45	32		13	92	88	43	38		7
60	107	73	13	44		16	101	79	19
90	103		13	115		25	50		40
120	148		28	62	118	2	91		29
150	54	126	24	115		35	71	109	41
180	181		1	56	124	56	113		67
225	279	261	36	319	221	4	295	245	20
270	240		30	342		72	316		46
300	207	333	33	241	299	1	306		6
330	253	287	43	333		3	329		1
345	326		19	333		12	347		2

Azimuth	2nd *On*	*Rev*	*Error*	2nd *Off*	*Rev*	*Error*	3rd *On*	*Rev*	*Error*
0	14		14	24		24	356		4
30	65		35	38		8	77		47
45	46		1	117	63	18	151	29	16
60	90		30	89		29	25		35
90	120		30	109		19	55		35
120	70	110	10	133		13	161		41
150	72	108	42	70	110	40	90		60
180	7	173	7	13	167	13	34	146	34
225	270		45	235		10	307	233	8
270	257		13	307		37	316		46
300	321		21	296		4	336		36
330	272		58	208	332	2	321		9
345	328		17	40		55	292		53

Participant 7

Azimuth	*Direct*	*Rev*	*Error*	1st *On*	*Rev*	*Error*	1st *Off*	*Rev*	*Error*
0	180	0	0	0		0	0		0
30	130	50	20	19		11	0		30
45	144	36	9	58		13	23		22
60	121	59	1	55		5	55		5
90	114		24	144		54	29		61
120	137		17	110		10	39	141	21
150	149		1	31	149	1	75	105	45
180	181		1	0	180	0	0	180	0
225	229		4	295	245	20	271	269	44
270	284		14	333		63	298		28
300	243	297	3	298		2	256	284	16
330	210	330	0	343		13	220	320	10
345	220	320	25	273		72	334		11

Azimuth	2nd *On*	*Rev*	*Error*	2nd *Off*	*Rev*	*Error*	3rd *On*	*Rev*	*Error*
0	0		0	329		31	0		0
30	197	-17	47	28		2	25		5
45	57		12	57		12	30		15
60	109	71	11	113	67	7	127	53	7
90	63		27	109		19	36		54
120	112		8	113		7	141		21
150	71	109	41	21	159	9	35	145	5
180	0	180	0	0	180	0	0	180	0
225	270		45	271	269	44	273	267	42
270	269		1	231		39	259		11
300	317		17	239	301	1	241	299	1
330	309		21	305		25	293		37
345	338		7	311		34	314		31

Participant 8

Azimuth	*Direct*	*Rev*	*Error*	1^{st} *On*	*Rev*	*Error*	1^{st} *Off*	*Rev*	*Error*
0	0		0	0		0	352	-8	8
30	63		33	33		3	162	18	12
45	118	62	17	78		33	105	75	30
60	77		17	90		30	98	82	22
90	85		5	99		9	102		12
120	101		19	121		1	137		17
150	163		13	136		14	147		3
180	180		0	180		0	181		1
225	263		38	259		34	271	269	44
270	270		0	282		12	261		9
300	263	277	23	298		2	280		20
330	290		40	293		37	284		46
345	340		5	334		11	322		23

Azimuth	2^{nd} *On*	*Rev*	*Error*	2^{nd} *Off*	*Rev*	*Error*	3^{rd} *On*	*Rev*	*Error*
0	0		0	0		0	0		0
30	126	54	24	159	21	9	127	53	23
45	103	77	32	111	69	24	136	44	1
60	100	60	0	106	74	14	155	25	35
90	70		20	97		7	91		1
120	90		30	90		30	91		29
150	148		2	159		9	151		1
180	359	181	1	180		0	180		0
225	253		28	262		37	270		45
270	291		21	283		13	265		5
300	271		29	278		22	241	299	1
330	321		9	282		48	297		33
345	337		8	346		1	283		62

Participant 9

Azimuth	*Direct*	*Rev*	*Error*	1^{st} *On*	*Rev*	*Error*	1^{st} *Off*	*Rev*	*Error*
0	181	-1	1	8		8	-5		5
30	105	75	45	44		14	71		41
45	89		44	37		8	77		32
60	11	66	6	113	67	7	53		7
90	91		1	59		31	78		12
120	145		25	66	114	6	85	95	25
150	149		v1	118		32	134		16
180	182		2	180		0	339	201	21
225	218		7	277	263	38	309	231	6
270	237		33	294		24	295		25
300	270		30	282		18	251	289	11
330	345		15	231	309	21	347		17
345	257	283	62	328		17	330		15

Azimuth	2^{nd} *On*	*Rev*	*Error*	2^{nd} *Off*	*Rev*	*Error*	3^{rd} *On*	*Rev*	*Error*
0	8		8	-8		8	-10		10
30	40		10	61		31	17		13
45	46		1	55		10	42		3
60	109	71	11	122	58	2	73		13
90	122		32	72		18	72		18
120	91		29	104		16	121		1
150	83	97	53	24	156	6	73	107	43
180	0	180	0	0	180	0	0	180	0
225	252		27	225		0	256		31
270	278		8	269		1	247		23
300	248	292	8	259	281	19	209	331	31
330	326		4	288		42	190	350	20
345	361		16	340		5	338		7

Participant 10

Azimuth	*Direct*	*Rev*	*Error*	1^{st} *On*	*Rev*	*Error*	1^{st} *Off*	*Rev*	*Error*
0	0		0	180	0	0	11		11
30	62		32	50		20	99	81	51
45	54		9	105	75	30	47		2
60	106	74	14	90		30	89		29
90	90		0	87		3	90		0
120	111		9	91		29	91		29
150	138		12	116		34	56	124	26
180	182		2	180		0	0	180	0
225	226		1	245		20	323	217	8
270	286		16	303		33	289		19
300	269	271	29	292		8	309		9
330	334		4	337		7	308		22
345	239	301	44	184	356	11	352		7

Azimuth	2^{nd} *On*	*Rev*	*Error*	2^{nd} *Off*	*Rev*	*Error*	3^{rd} *On*	*Rev*	*Error*
0	166	14	14	181	-1	1	0		0
30	39		9	103	77	47	14		16
45	72		27	87		42	55		10
60	118		58	80		20	110	70	10
90	88		2	90		0	81		9
120	90		30	118		2	109		11
150	102		48	128		22	117		33
180	180		0	0	180	0	0	180	0
225	228		3	305	235	10	244		19
270	303		33	311		41	292		22
300	317		17	290		10	288		12
330	313		17	321		9	233	307	23
345	115		30	309		36	303		42

Participant 11

Azimuth	*Direct*	*Rev*	*Error*	1^{st} *On*	*Rev*	*Error*	1^{st} *Off*	*Rev*	*Error*
0	201	-21	21	346		14	3		3
30	50		20	6		24	41		11
45	109	71	26	58		13	21		24
60	131	49	11	83		23	131	49	11
90	103		13	44		46	109		19
120	150		30	24	156	36	65	115	5
150	166		16	119		31	109		41
180	183		3	356	184	4	359	181	1
225	212		13	326	214	11	280	260	35
270	226		44	285		15	297		27
300	237	303	3	352		52	299		1
330	218	322	8	340		10	309		21
345	348		3	346		1	342		3

Azimuth	2^{nd} *On*	*Rev*	*Error*	2^{nd} *Off*	*Rev*	*Error*	3^{rd} *On*	*Rev*	*Error*
0	178	2	2	183	-3	3	349	-11	11
30	138	42	12	161	19	11	144	36	6
45	31		14	100	80	35	148	32	13
60	76		16	119	61	1	146	34	26
90	107		17	116		26	61		29
120	45	135	15	132		12	68	112	8
150	16	164	14	6	174	24	40	140	10
180	342	198	18	0	180	0	0	180	0
225	301	239	14	250		25	266		41
270	272		2	244		26	237		33
300	266	274	26	249	291	9	207	333	33
330	262	278	52	329		1	279		51
345	319		26	218	322	23	221	319	26

Participant 12									
Azimuth	*Direct*	*Rev*	*Error*	1^{st} *On*	*Rev*	*Error*	1^{st} *Off*	*Rev*	*Error*
0	215	35	35	0		0	180	0	0
30	108	72	42	71		41	165	15	15
45	100	80	35	100	80	35	52		7
60	91	89	29	101	79	19	100	80	20
90	80		10	72		18	77		13
120	77	103	17	106		14	102		18
150	131		19	103		47	166		16
180	144		36	180		0	180		0
225	260		35	263		38	289	251	26
270	260		10	288		18	270		0
300	265	275	25	270		30	278		22
330	243	297	33	237	303	27	310		20
345	286		59	217	323	22	180	360	15
Azimuth	2^{nd} *On*	*Rev*	*Error*	2^{nd} *Off*	*Rev*	*Error*	3^{rd} *On*	*Rev*	*Error*
0	180	0	0	0		0	0		0
30	132	48	18	49		19	70		40
45	126	54	9	49		4	114	66	21
60	77		17	91		31	102	78	18
90	87		3	89		1	82		8
120	102		18	90		30	93		27
150	77	103	47	104		46	53	127	23
180	180		0	0	180	0	180		0
225	286	254	29	282	258	33	279	261	36
270	270		0	257		13	253		17
300	253	287	13	264	276	24	258	282	18
330	251	289	41	308		22	257	283	47
345	209	331	14	321		24	232	308	37

Table A.1: Results from the virtual Ambisonics localisation listening tests.

Appendix B

Room Impulse Response Database

Convolution reverberation has become increasingly popular as more impulse responses (IRs) are becoming available along with more advanced tools to manipulate those IRs. IRs also give insight into how sound waves travel through a room allowing researchers to define and refine models to better predict how the acoustics of a space will sound and why.

If a room is considered to be an LTI system, then taking a measurement of that system allows predictions and simulations to be made. When the system being examined is a room, an IR is measured by recording sounds like a pistol shot or popped balloon and the subsequent reverberation. By convolving that recording with a recording of a sound source without reverberation, the result sounds as if that sound source was originally recorded in the room that was measured.

Convolution can create a far more natural sounding artificial reverberation than other techniques, but it has its limitations. Specifically, convolution reverberation is limited by the IRs that are available. A single IR can only accurately represent the exact sound source and receiver configuration that were used during the measurement. A single IR cannot create a moving sound source or receiver position and it does not provide complete information about measurements from other locations in the same room.

B.1 Other Available Databases

A number of databases of IRs are currently available. The Aachen Impulse Response (AIR) database [Jeub et al., 2009] is explicitly developed for hearing aid research. It uses a dummy head to create head shadowing effects in the IRs. By introducing these effects, the IRs are considered binaural room impulse response (BRIRs). A total of 64 IRs (without a dummy head) and BRIRS (with a dummy head) are measured across four different rooms ranging from 12 m^3 to 370 m^3. The measurement technique used is a maximum length sequence; see [Schroeder, 1965, Holters et al., 2009]. There are a maximum number of six different source and receiver configurations in a single room.

The Multichannel Acoustics Reverberation Database at York (MARDY) is a collection of multi-channel IRs recorded in the recording studio at the University of York [Wen et al., 2006]. Eight multi-channel IRs are recorded in the same studio with two different acoustic panels and four different source and receiver configurations.

Murphy recorded a third collection of IRs measured in a series of sites with significant archaeological acoustics [Murphy, 2006] using a setup designed by Farina and Ayalon [2003]. The setup consisted of a cardioid microphone and a first order Ambisonic B-format microphone. The IRs from the cardioid microphone can be combined with other IRs to create a stereo pair. The two microphones were mounted on a rotating turntable so that they automatically moved and took a total of 72 measurements across 5° intervals of a full circle. The IRs were measured with a logarithmic sine sweep [Farina, 2000, Holters et al., 2009].

Waves Audio Ltd. uses the rotating turntable setup also used in [Murphy, 2006] but includes a binaural dummy head to create BRIRs [Ben-Hador and Neoran, 2004] as done in [Farina and Ayalon, 2003]. The IRs are mixed to various output formats using the measured IRs and virtual microphones. They are available online[1] and intended for use with the Waves IR products. The IR library is intended for music production and includes over 100 different spaces with up to three source and receiver configurations for stereo and surround IRs.

All of these databases are limited in some way due to to physical constraints, whether by the microphones or the source and receiver configurations. In particular, they do not provide a dense grid of measurements across a single room. The collection described in [Murphy, 2006] does measure 72 locations in the same room, but those 72 measurements are very near each other especially relative to the size of the spaces being measured. Acoustics research has repeatedly reported that acoustic metrics can vary greatly across a single space (see [Bradley, 1994, de Vries et al., 2001] for a sample of these findings), but the currently available databases do not provide sufficient real-world data for analysis. Many researchers record their own measurements for their research, but do not release those measurements for others to use. A publicly-available database that broadly samples across a room is needed. The database presented here provides an order of magnitude more measurements and covers a significantly greater percentage of the floor area than previous publicly released databases.

B.2 Measurement Technique

All three sets of IRs were measured in spaces at Queen Mary, University of London, London, UK. The rooms have different acoustic properties, but all have significant reverberation times. All IRs were measured with a Genelec 8250A loudspeaker as the source and each receiver position was measured with an omnidirectional DPA 4006 and a B-format Soundfield SPS422B. The loudspeaker and Soundfield microphone can be seen in Figure B.1.

It is a less complicated process to simulate binaural measurements from omnidirectional IRs than it is to remove binaural information from binaural IRs. While the Aachen Impulse Response database was intentionally designed for applications that prefer binaural information, our database is designed to be as flexible as possible. The physical labor required to record a set of measurements in a single room is prohibitively time-consuming, so only two different microphones were selected. The sound source was

[1] http://acoustics.net

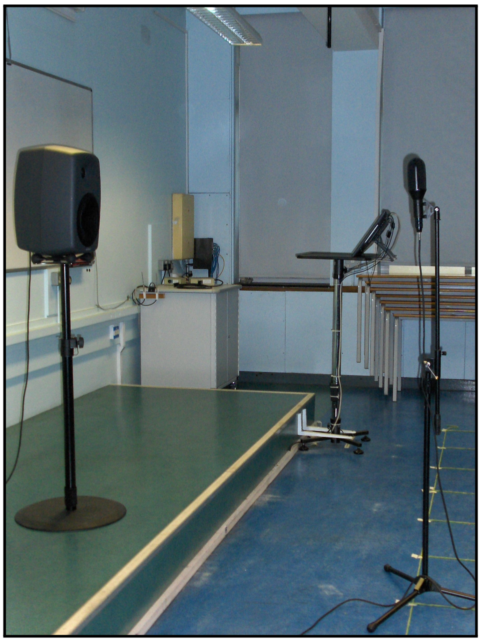


Figure B.1: The Soundfield microphone and Genelec speaker during measurements in the classroom.

kept static and the receivers moved around the space, as is common in IR databases. This procedure was well suited to the study as two of the three rooms measured have a clear single location for the sound source (lecturing platform or stage) with multiple listening locations.

Omnidirectional IRs were recorded with a DPA 4006 and with the W-channel of the Soundfield microphone. The Soundfield microphone does not have a flat frequency response, particularly at higher frequencies whereas the DPA 4006 has a much flatter response. The more accurate capture of the high frequencies may be preferred for some applications.

The IRs were measured using the logarithmic sine sweep described in [Farina, 2000]. This approach has been shown to minimize harmonic distortion and increase the signal-to-noise ratio without needing to average multiple measurements, see [Holters et al., 2009].

B.3 Database of Impulse Responses

Each measurement had source and receiver heights of 1.5 m.

B.3.1 Classroom

A set of 130 IRs were taken within a classroom in the School of Electronic Engineering and Computer Science. The room measures roughly 7.5 x 9 x 3.5 m (236 m^3) with reflective surfaces of a linoleum floor, painted plaster walls and ceiling, and a large whiteboard. When in use for lectures the room is filled with desks and chairs. These were stacked and moved to the side against the windows during the measurements. Measurements were 500 cm apart arranged in 10 rows and 13 columns relative to the speaker, with the 8th column directly on axis with the speaker. Figure B.1 shows the classroom and the

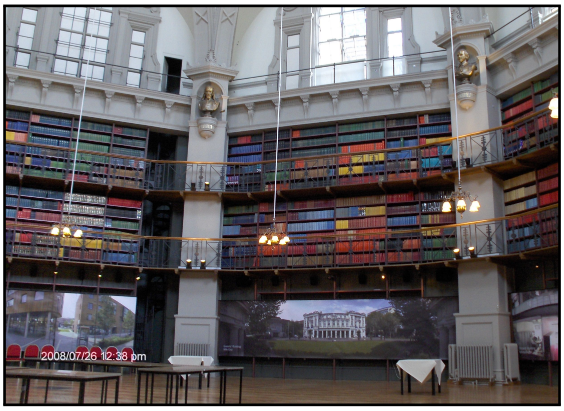


Figure B.2: The Octagon.

dimensions of the room and the receiver positions can be seen in Figure B.3.

B.3.2 Octagon

The Octagon is a Victorian building at Queen Mary completed in 1888 and originally designed to be a library. It is currently used as a conference venue, but the walls are still lined with books with a wooden floor and plaster ceiling. As the name suggests, the room has eight walls each 7.5 m in length and a domed ceiling reaching 21 m over the floor, with an approximate volume of 9500 m^3. A set of 169 IRs were measured in the centre of the room. A diagram of the room and the speaker and microphone positions can be seen in Figure B.3 and a photo of the room in Figure B.2.

B.3.3 Great Hall

The Great Hall is a multipurpose hall that can hold approximately 800 seats. The hall has a stage and seating areas on the floor and a balcony. The microphones were placed in the seating area on the floor, roughly a 23 m x 16 m area, which was cleared of chairs. The microphone positions were identical to the layout for the Octagon, 169 IRs over a 12 m x 12 m area. Figure B.3 shows the area where the IRs were measured, but the room is significantly bigger as the balcony extends 20 m past the rear wall.

B.4 Availability

The C4DM RIR database is available for download from `http://isophonics.net/content/room-impulse-responses`. All IRs are mono 96 kHz, 32 bit wav files including the B-format IRs, so four files are required for a full three-dimensional B-format IR. The database is released under the Creative Commons Attribution-Noncommercial-Share Alike license with attribution to the Centre for Digital Music, Queen Mary, University of London.

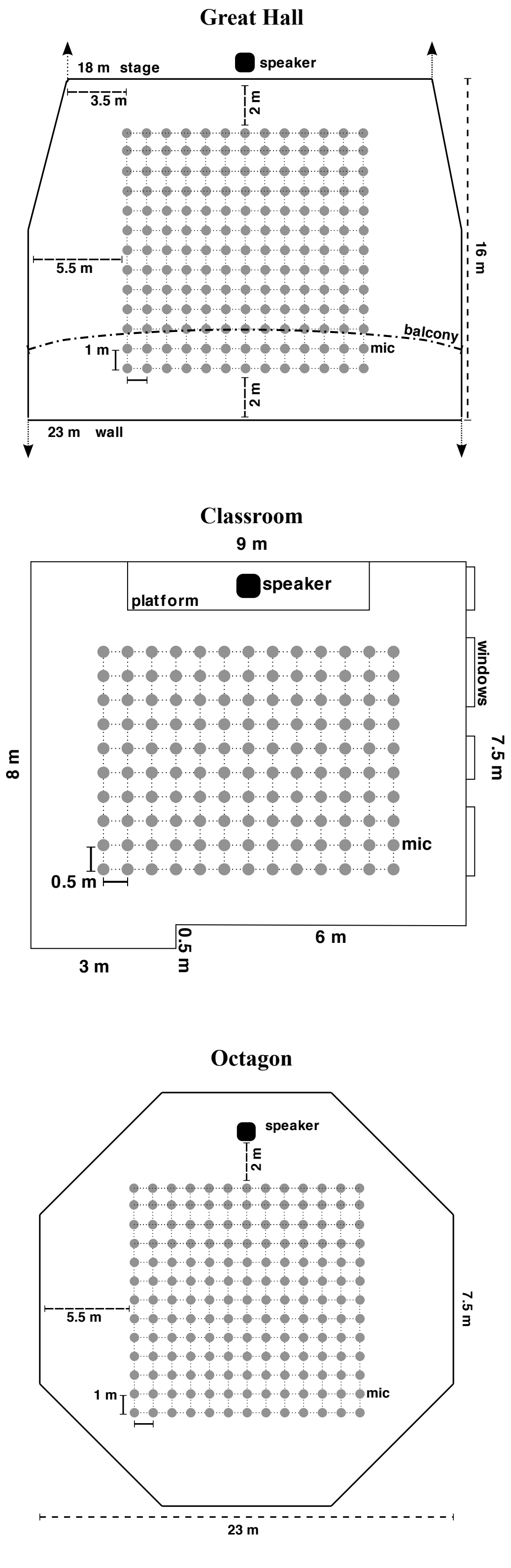


Figure B.3: Source and receiver positions of the three rooms.

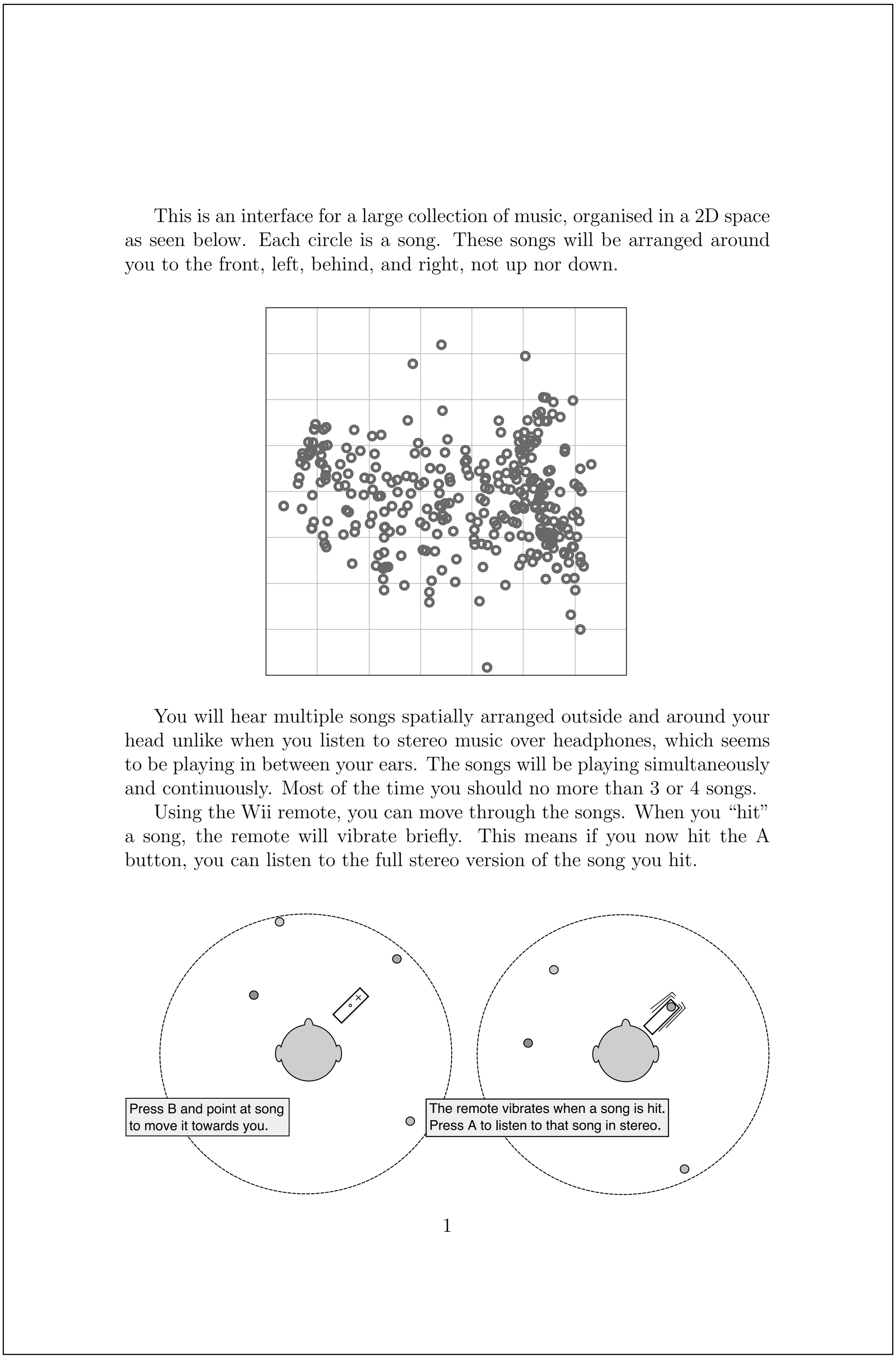
Appendix C

amblr User Evaluation

This appendix includes the raw data gathered and instructions and survey used for the user evaluation in Section 6.1.4. Below is the table of the results from the survey. Figures C.1 through C.3 show the instructions given to each participant, and Figures C.4 and C.5 shows the survey.

Participant	*Gender*	*Age*	*Travel*	*Localise*	*Approach*	*Zoom*
1	male	25-30	2	6	2	2
2	male	31-35	2	5	3	6
3	male	25-30	3	3	4	7
4	male	25-30	2	2	4	3
5	female	25-30	5	4	6	4
6	female	31-35	5.5	6	6	4
7	male	36+	5	5	3	5
8	female	25-30	3	4	3	4
9	male	36+	6	7	6	4
10	male	25-30	5	3	4	3
11	male	25-30	2	5	3	4
12	female	26+	5	2	2	6

Table C.1: Results gathered from the amblr user evaluation.



This is an interface for a large collection of music, organised in a 2D space as seen below. Each circle is a song. These songs will be arranged around you to the front, left, behind, and right, not up nor down.

You will hear multiple songs spatially arranged outside and around your head unlike when you listen to stereo music over headphones, which seems to be playing in between your ears. The songs will be playing simultaneously and continuously. Most of the time you should no more than 3 or 4 songs.

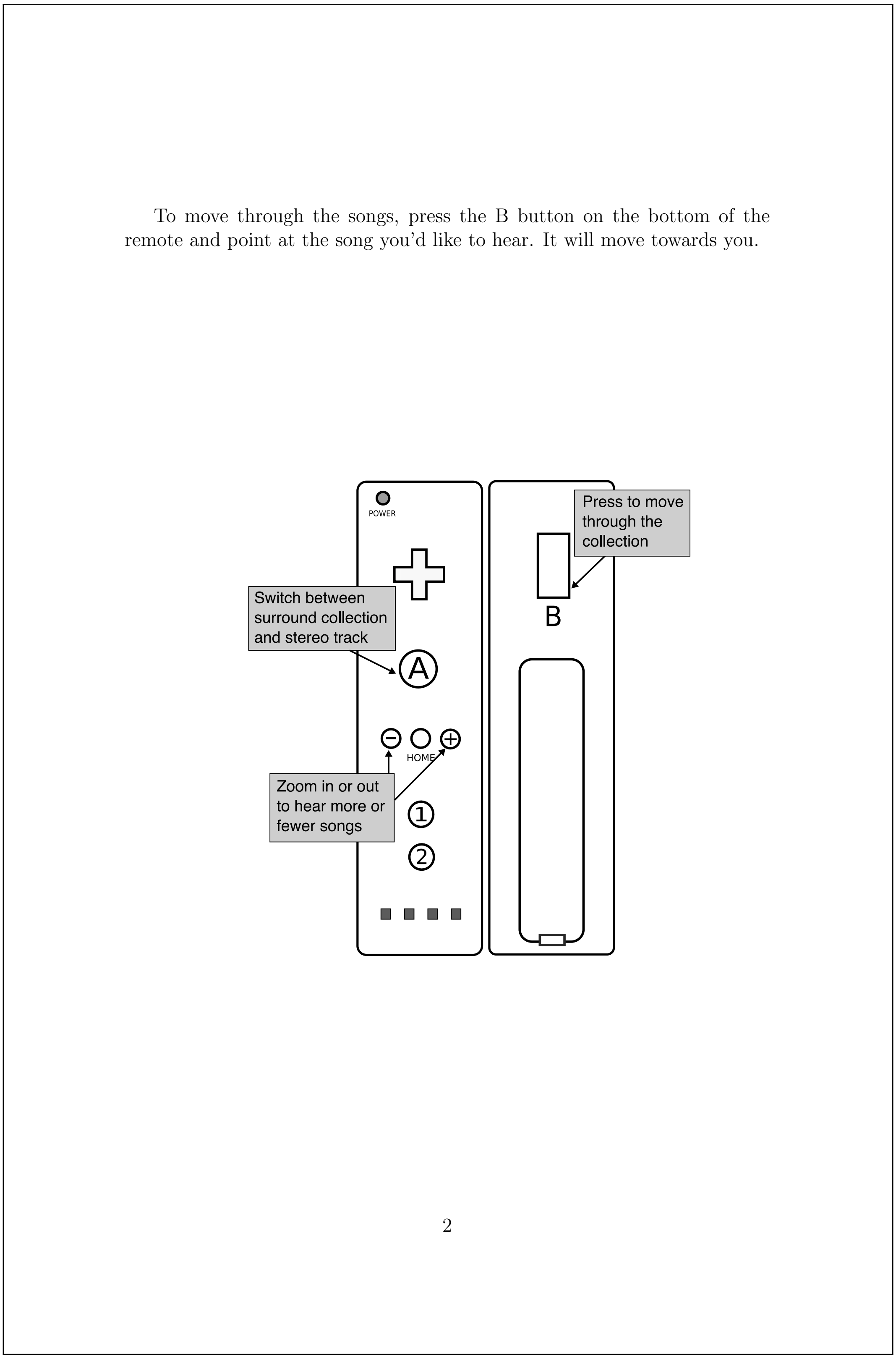
Using the Wii remote, you can move through the songs. When you "hit" a song, the remote will vibrate briefly. This means if you now hit the A button, you can listen to the full stereo version of the song you hit.

Press B and point at song to move it towards you.

The remote vibrates when a song is hit. Press A to listen to that song in stereo.

1

Figure C.1: First page of the instructions given to user evaluation participants.

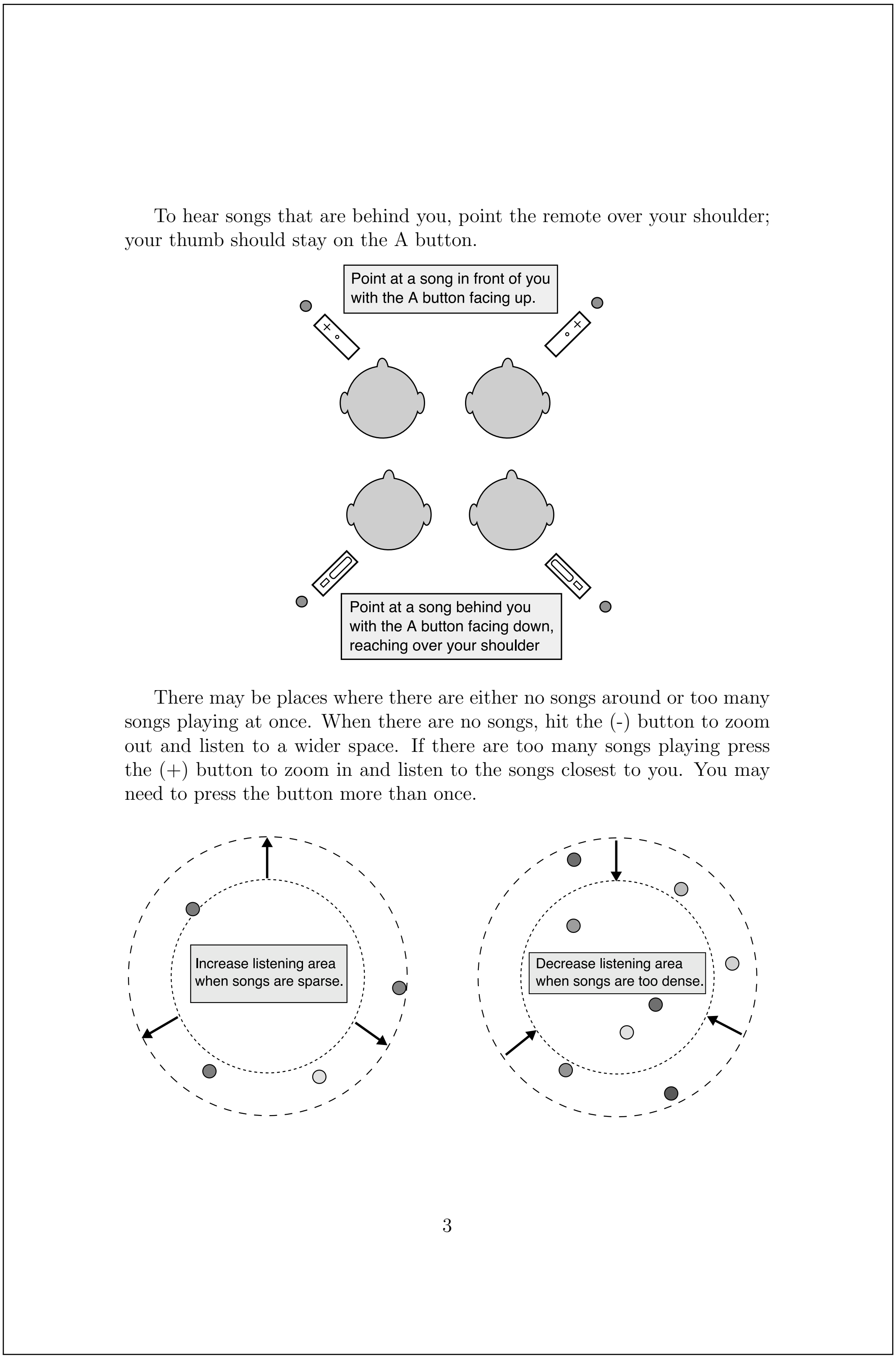


To move through the songs, press the B button on the bottom of the remote and point at the song you'd like to hear. It will move towards you.

POWER
Press to move through the collection
B
Switch between surround collection and stereo track
A
HOME
Zoom in or out to hear more or fewer songs
1
2

2

Figure C.2: Second page of the instructions given to user evaluation participants.



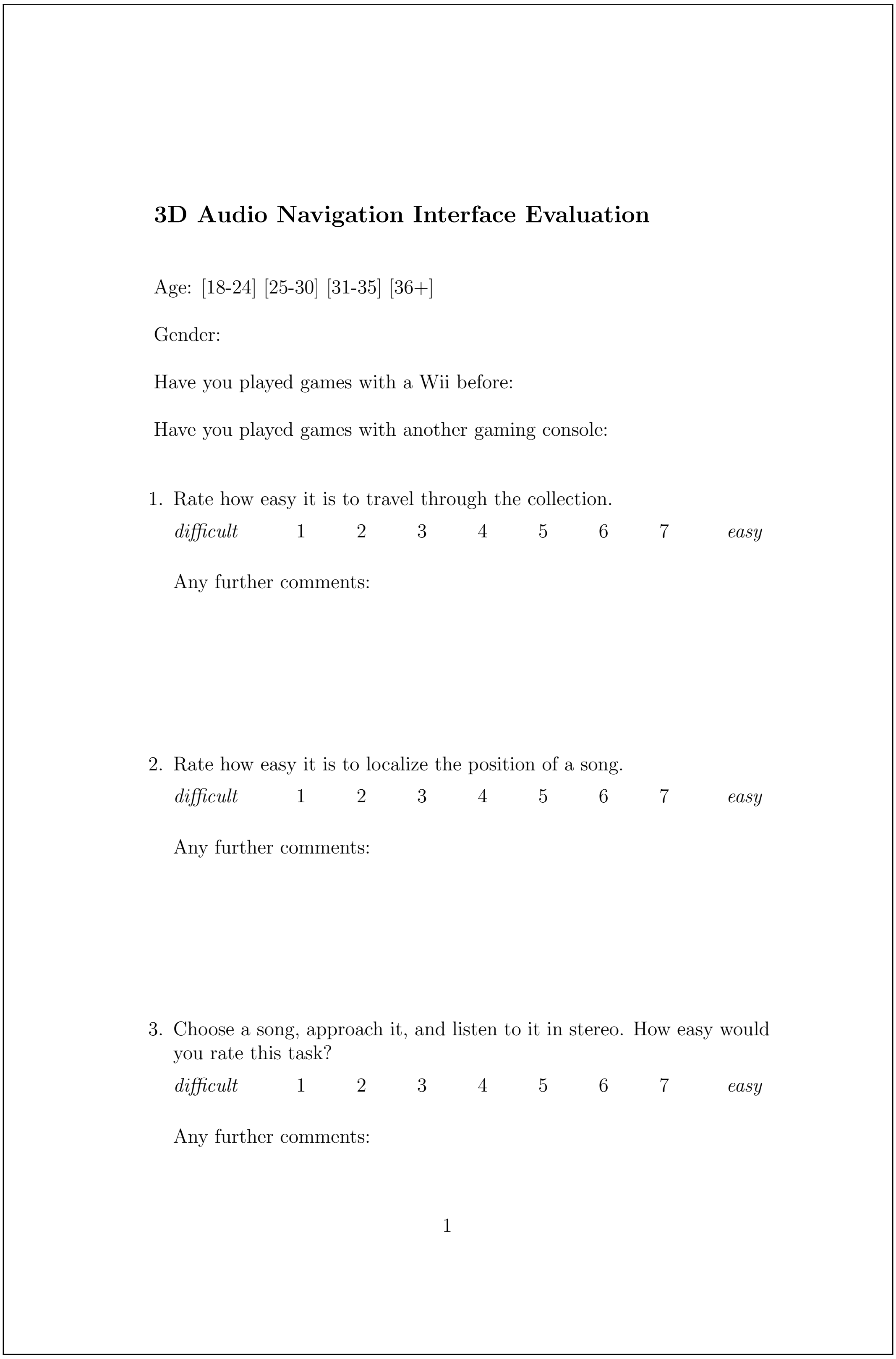
To hear songs that are behind you, point the remote over your shoulder; your thumb should stay on the A button.

Point at a song in front of you with the A button facing up.
Point at a song behind you with the A button facing down, reaching over your shoulder

There may be places where there are either no songs around or too many songs playing at once. When there are no songs, hit the (-) button to zoom out and listen to a wider space. If there are too many songs playing press the (+) button to zoom in and listen to the songs closest to you. You may need to press the button more than once.

Increase listening area when songs are sparse.
Decrease listening area when songs are too dense.

Figure C.3: Third page of the instructions given to user evaluation participants.



3D Audio Navigation Interface Evaluation

Age: [18-24] [25-30] [31-35] [36+]

Gender:

Have you played games with a Wii before:

Have you played games with another gaming console:

1. Rate how easy it is to travel through the collection.
   difficult 1 2 3 4 5 6 7 *easy*

   Any further comments:

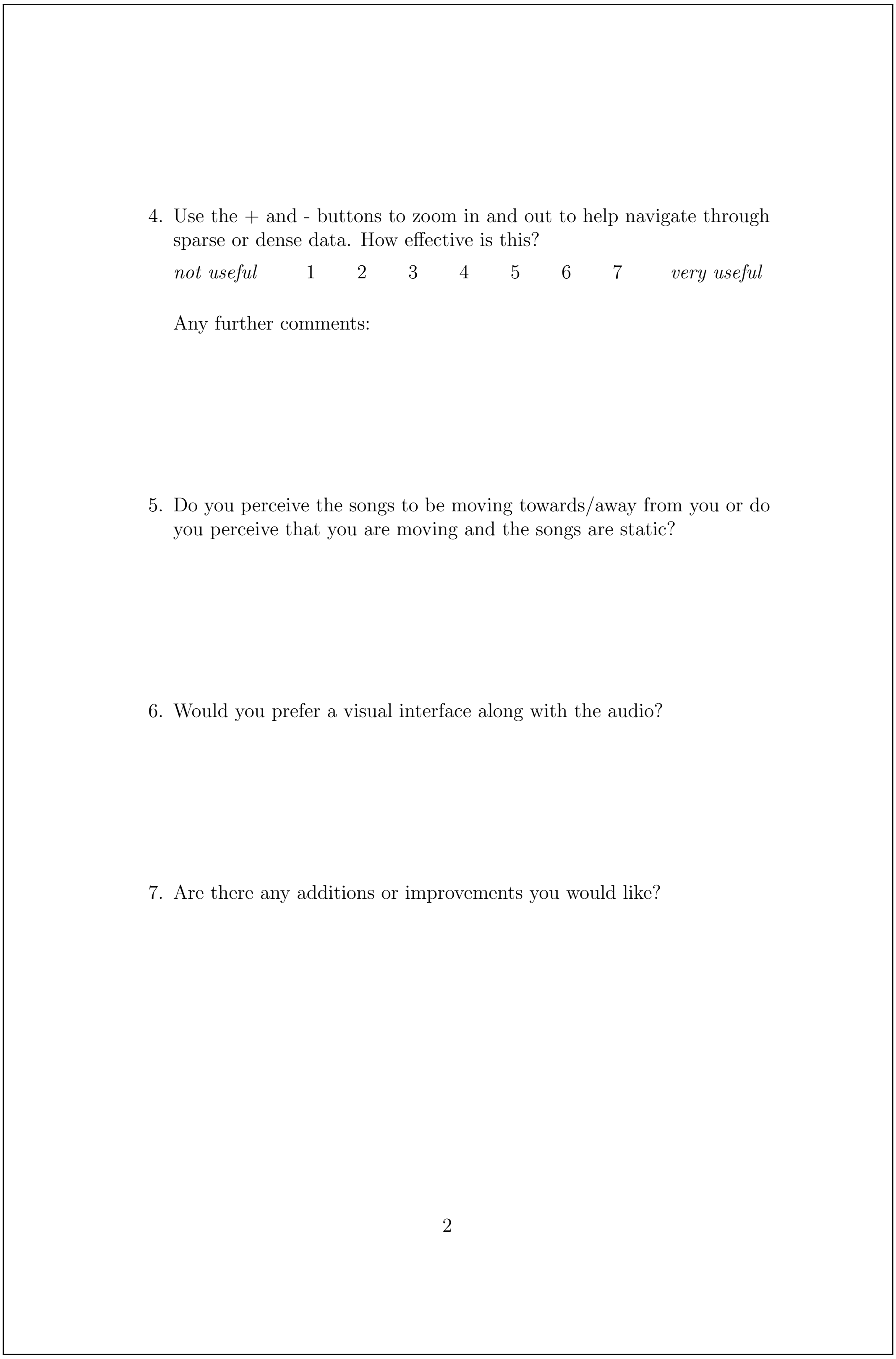
2. Rate how easy it is to localize the position of a song.
   difficult 1 2 3 4 5 6 7 *easy*

   Any further comments:

3. Choose a song, approach it, and listen to it in stereo. How easy would you rate this task?
   difficult 1 2 3 4 5 6 7 *easy*

   Any further comments:

Figure C.4: First page of the survey given to user evaluation participants.



4. Use the + and - buttons to zoom in and out to help navigate through sparse or dense data. How effective is this?

   not useful 1 2 3 4 5 6 7 *very useful*

   Any further comments:

5. Do you perceive the songs to be moving towards/away from you or do you perceive that you are moving and the songs are static?

6. Would you prefer a visual interface along with the audio?

7. Are there any additions or improvements you would like?

2

Figure C.5: Second page of the survey given to user evaluation participants.