A Comparative Study on the Perceived Sensation of Motion in Real and Virtual Environments

Matthew Boerum

Department of Music Research

Schulich School of Music

McGill University, Montreal

2020

A thesis submitted to McGill University in partial fulfilment

of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

This thesis consists of a three-part study to investigate the *sensation of motion* as perceived by listeners during *three-dimensional* (3D) audio production scenarios exhibiting auditory motion. The study and experimental methodology are based on previous research in binaural audio simulation, psychoacoustics, 3D audio, *virtual reality* (VR) and auditory motion. Experiments are designed to produce auditory motion using established 3D audio panning techniques. The study then conducts three sequential experiments within an anechoic, reverberant and virtual reality environment. Perceptual evaluations show there is no substantial environmental influence on the *sensation of motion* when all aspects of the real environment are replicated within the virtual environment. Visual representations of the virtual environments are then presented within the experiments, which show a significant influence on the sensation of motion as a whole, however, environmental influence remains insignificant. Further analysis illustrates the influence of reverberation and lateralization on the sensation of motion as a whole. The final experiment performed within a real and replicated virtual environment, through the use of VR, demonstrates that the presence of virtual reality produced no significant effect on the ability to perform a 3D audio panning task which produces the sensation of motion. This research was conducted to determine correlations between reality and virtual reality such that audio engineers are better informed on how the perception of auditory motion in virtual environments translates to a sensation of motion for the user of VR and *augmented reality* (AR). Little to no evidence was found in the experimental results that would suggest that binaural crossfading, nor the presence of

replicated virtual environments, should inhibit the sensation of motion as a comparative perceptual rating for simulated and referenced auditory motion. Limited or restricted visual cues, and changes in acoustics have a larger influence on the sensation of motion than the virtual environment itself.

RÉSUMÉ

Cette thèse consiste en une étude en trois parties pour étudier la sensation de mouvement telle que perçue par les auditeurs lors de scénarios de production audio en trois dimensions (3D) présentant un mouvement auditif. L'étude et la méthodologie expérimentale sont basées sur des recherches antérieures en simulation audio binaurale, psychoacoustique, audio 3D, réalité virtuelle (VR) et mouvement auditif. Les expériences sont conçues pour produire un mouvement auditif en utilisant des techniques de panoramique audio 3D établies. L'étude mène ensuite trois expériences séquentielles dans un environnement de réalité anéchoïque, réverbérante et virtuelle. Les évaluations perceptives montrent qu'il n'y a pas d'influence environnementale substantielle sur la sensation de mouvement lorsque tous les aspects de l'environnement réel sont reproduits dans l'environnement virtuel. Des représentations visuelles des environnements virtuels sont ensuite présentées au sein des expériences, qui montrent une influence significative sur la sensation de mouvement dans son ensemble, cependant, l'influence environnementale reste insignifiante. Une analyse plus approfondie illustre l'influence de la réverbération et de la latéralisation sur la sensation de mouvement dans son ensemble. L'expérience finale réalisée dans un environnement virtuel réel et répliqué, grâce à l'utilisation de la VR, démontre que la présence de la réalité virtuelle n'a produit aucun effet significatif sur la capacité à effectuer une tâche de panoramique audio 3D qui produit la sensation de mouvement. Cette recherche a été menée pour déterminer les corrélations entre la réalité et la réalité virtuelle afin que les ingénieurs du son soient mieux informés sur la façon dont la perception du mouvement auditif dans les environnements virtuels

se traduit par une sensation de mouvement pour l'utilisateur de la réalité virtuelle et de la réalité augmentée (AR). Peu ou pas de preuves ont été trouvées dans les résultats expérimentaux qui suggéreraient que le crossfading binaural, ni la présence d'environnements virtuels répliqués, devraient inhiber la sensation de mouvement comme évaluation perceptive comparative pour le mouvement auditif simulé et référencé. Les indices visuels limités ou restreints et les changements d'acoustique ont une plus grande influence sur la sensation de mouvement que l'environnement virtuel lui-même.

ACKNOWLEDGEMENTS

This research was supported by the Sound Recording Program in the Schulich School of Music at McGill University along with the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT). I would like to thank the Schulich Family, Schulich School of Music, McGill University Alumni Association, CIRMMT, Mitacs Accelerate program, LANDR, and PlayHybrid for the scholarships and other financial support which allowed me to attend McGill University and complete this research.

My thesis is advised by Richard King and George Massenburg. I would like to express my deepest gratitude to them, and to Wieslaw Woszczyk and Martha de Francisco for providing me with an opportunity to greatly challenge myself and reach a level of expertise that I never knew was possible. In addition, I would like to thank Richard King and George Massenburg for your incredible skill and expertise, along with your unwavering support and willingness to help.

I would also like to thank my PhD colleagues (Dave Benson, Will Howie, Jack Kelly, Bryan Martin, Denis Martin, and Diego Quiroz) for all of your help along the way, whether it was teaching me MATLAB, helping me build my experiments, lending a hand in listening tests, or just unwinding at Thompson House. I couldn't have done it without you! In addition, I would like to acknowledge and thank Yves Méthot and Julien Boissinot of CIRMMT, Jerry Catanescu, Ilja Frissen, Patrick Cheiban, Lasse Vetter, Michael Terrell, Adrien Quillet, and Mike Greenfield for supporting the design of my experiments.

Finally, I would like to thank my amazing, loving family for their endless encouragement and support during this process. To my two little boys who put up with work-filled evenings and weekends for far too long—thank you! Dad can finally play! Last but certainly not least, this degree would have never happened without the selfless support and encouragement of my best friend and wife, Mary Catherine who agreed to temporarily postpone both of our professional careers, pack up our family for Montréal and bet everything on McGill and this degree. It was absolutely worth it. Thank you—we celebrate this together!

PREFACE

This thesis is presented in manuscript form. The content and experiments from Chapters 3, 4, and 5 were previously published by the Audio Engineering Society over the course of 2½ years. Most of the content (including text, equations, figures, tables, and references) appears as originally drafted. However, due to the previous publication length restrictions of the articles presented in Chapters 3, 4 and 5, additional information has now been provided to present greater detail on methodologies, analyses and conclusions. Additional figures and support for the topics presented in these chapters are included in the Chapter 6, Appendix A and Appendix B. Additionally, updates and formatting were necessary to align the points of each of these articles into one cohesive document. The figures and images in this thesis are the original work of Matthew Boerum, unless otherwise stated. The co-authors of the previously published articles have provided written permission to reproduce these articles as part of the manuscript thesis.

ORIGINAL SCHOLARSHIP AND DISTINCT CONTRIBUTIONS TO KNOWLEDGE

Many studies on the perception of auditory motion have focused on the perceptual accuracy and awareness of auditory cues during motion events by evaluating, for example, minimum audible movement angle, factors of auditory vection, or the accuracy and efficiency of head-related transfer function (HRTF) databases. However, the author feels that it would serve the audio engineering industry well to develop further research on the quality of these auditory events in addition to accuracy and awareness, with the hope of understanding the translational factors necessary to control quality during 3D audio productions. This thesis investigates an area of research in which perceptual factors of virtual simulations are studied and compared to real-world events using production-based techniques like panning. The work in this thesis offers novel approaches to study such translational percepts and is the first investigation to directly investigate the sensation of motion percept as a variable of virtual environment quality and influence.

Chapters 3 and 4 detail experimental design and perceptual evaluation methods that are novel solutions to the comparison of auditory motion perception in real and virtual environments. Most existing methodologies are designed to conduct this experiment using real human movement, which is unreliable across subjects, or through virtual simulations compared against predictable models. To ensure real-world auditory motion be consistently captured and referenced across multiple subjects, a translational (longitudinal) motion apparatus was constructed for repeatable

measurements. This apparatus was designed to eliminate noise and vibration, while providing a seamless motion action that could be replicated through video tracking, as well as through an audio reference. It is the first study of its kind to capture real-world binaural recordings during self-motion (using a dummy head) to directly compare the perception of auditory motion to a replicated virtual simulation. This method can be used to conduct further experiments on the translational aspects of motion perception for active listeners in real and virtual environments. The results of this two-part study show that virtual environment design should accurately replicate the real-world reference to maintain quality in translation. It also revealed that binaural crossfading is a simulation technique that provides equal or increased sensation of motion in virtual environments. The method used for the perceptual evaluation of the sensation of motion was a novel approach which combined quality and motion perception percepts into one listening experiment. This evaluation method can serve to influence further development of motion quality evaluation methods. Finally, the modular design of the motion apparatus allowed for the experiment to be replicated in a reflective environment to study the influence of reflections on the sensation of motion. These results revealed the surprising advantage of binaural crossfading to reduce the effects of binaural masking.

Chapter 5 introduced an application-based experiment that served to provide insight on virtual reality as a new medium for audio production when the listener requires accurate translation of the sensation of motion. The design of the experiment was novel in that it was the first experiment to replicate a hemispherical speaker array in a virtual environment which was presented to the user

in VR while wearing a *head-mounted display* (HMD). It also used novel methods for controlling auditory motion in a non-biased application (through software and hardware). The experimental method completely immersed subjects into a virtual environment in which audiovisual cues were perfectly translated from the real world. The modular design of the virtual elements provided full control over experimental variables such as speaker placement, room features, and listening position. The results of this study were heavily dependent on this novel approach and confirmed the results of Chapter 3 and 4. The virtual reality environment assets have been donated to the Sound Recording Program at McGill University to support further studies in virtual reality.

CONTRIBUTIONS OF AUTHORS

For the three previously published articles reproduced within this thesis, I was the first author and primary contributor to the hypotheses, experimental designs, original concepts, methodologies, evaluation software, interpretation of results, and organization of listening test participants.

Chapter 3

M. Boerum, B. Martin, R. King, G. Massenburg, D. Benson, W. Howie, "Lateral Listener Movement on the Horizontal Plane: Sensing Motion Through Binaural Simulation," in AES 61st International Conference: Audio for Games, London, 2016.

Bryan Martin helped conduct the reference measurements and apparatus construction. Richard King provided guidance on the experimental direction as well as final editing of the written document. George Massenburg provided guidance on the experimental design. Dave Benson conducted the statistical analysis of the perceptual results data in R Studio. Will Howie assisted in administration of the perceptual listening tests.

Chapter 4

M. Boerum, B. Martin, R. King, G. Massenburg, "Lateral Listener Movement on the Horizontal Plane: Part 2 Sensing Motion Through Binaural Simulation in a Reverberant Environment," in AES Conference on Audio for Virtual and Augmented Reality, Los Angeles, 2016.

Bryan Martin assisted with the reference measurements and apparatus construction. Richard King provided guidance on the experimental direction and final editing of the written document. George Massenburg provided guidance on the experimental design.

Chapter 5

M. Boerum, J. Kelly, D. Quiroz, P. Cheiban, "The Effect of Virtual Environments on Localization during a 3D Audio Production Task," in AES Conference on Spatial Reproduction, Tokyo, 2018.

Jack Kelly supported the administration of the perceptual listening tests and collaborated on the experimental design. Diego Quiroz adapted the IRCAM VBAP panning system to the loudspeaker array, collaborated on the experimental design, and supported the administration of the perceptual listening tests. Patrick Cheiban assisted with statistical analysis of the perceptual results data.

CONTENTS

List of Tables

List of Equations

List of Figures

1 INTRODUCTION

For nearly a century, professional audio has been dependent on highly trained audio engineers with access to high quality loudspeakers and acoustically treated mixing studios to create commercial productions. This dependency on the highest quality ensures successful delivery of commercial multichannel mixes across a wide array of consumer playback sources. However, even for the best professionals, translating the mix from studio environment to the consumer product can be difficult since every audio playback source differs in quality and number of output channels. Even with these assets available to the professional, an educated guess must be made during the final mastering process since the relationship between studio reference and final consumer playback source is unpredictable, not only because of differences in playback environment, but also due to restrictions of playback hardware (digital audio converters, amplification, speakers, etc.).

However, through object-based audio, the multichannel studio reference is scaled proportionally to the playback source [1] [2] [3]. Object-based audio delivered through headphones completely removes any dependency on multichannel hardware as virtual speaker systems can be emulated through headphones using virtual head models. Object-based audio has therefore transformed the traditional mix translation process. Virtual reality benefits from object-based audio for its ability to present multiple simultaneous sound sources in 3D space. As a result, immersive audio for most

VR platforms is designed for headphones, and in most cases, produced on headphones. Engineers monitor their mixes within virtual reality to confirm audio objects are accurately connected to visual locations and that acoustic simulations match their visual counterparts.

Optimized, as it may seem, the VR platform still introduces many unpredictable variables that can alter the final presentation of the mix. Even in stereo productions, small changes in program material related to motion, such as sound source panning could completely change the experience of a production if inaccurately translated. In VR, motion is constant and often unpredictable since the user is in complete control of their experience at all times; this is the complete opposite of a traditional stereo mix.

The focus of mix translation for moving sound objects then becomes a matter of understanding the impact of the virtual environment on the perception of auditory motion. It also requires an understanding of the accuracy of common techniques for auditory motion simulation within the environment.

Since much of the existing literature for training professional audio engineers is structured for studio monitoring, we must look to understand the influence of virtual environment on the translation of audio mixes. For instance, a great sounding mixing studio can often lead to a better sounding final mix, and for this reason, engineers mix commercial films in sound stages which replicate the movie theatre environment. But this only simulates the environment where consumers will experience the film; it doesn't simulate the environment where the film actually takes place.

So, what is to be expected when engineers mix within the virtual environment using VR, and monitor this virtual space through headphones? Will mixes translate properly from a headphone-mixing environment to portray accurate virtual representations of the audible cues and acoustics in VR? Do visuals have an influence on immersive audio production at all? How do auditory motion events translate in VR? These questions remain unanswered and will be the focus of research for this dissertation.

1.1 Research Goals

The audio industry is rapidly evolving to adapt to new ways of consuming audio as virtual and augmented reality products are made more available. Virtual and augmented reality platforms present new commercial opportunities and challenges for audio engineers. While stereo and surround techniques within the studio are well-established and understood, with a vast amount of documented research and instructional documentation, techniques and methods for producing and mixing in VR and AR are much less established. In fact, content for the VR/AR audio industry is largely produced without the traditionally trained professional audio engineer. This presents a paradigm shift whereby the traditional audio engineer will need to adapt to new, platform-specific techniques.

The goal of this research is to determine correlations between reality and virtual reality such that audio engineers are better informed on how the perception of auditory motion translates to a sensation of motion for the user of VR and AR. It is also important for audio engineers to understand how a mix or production will translate from studio mixing environments to virtual

environments. By focusing on sound object motion, a key element used in both, stereo and 3D audio productions, this research will offer suggestions to help guide production techniques for the future. The experiments will also determine the dependency sound productions may have on visual accuracy within a virtual environment.

1.2 Structure of Thesis

This thesis is comprised of the following chapters:

Chapter 1 | Introduction is a summary of the inspiration, goals and objectives for this thesis.

Chapter 2 | Background is a detailed literature overview of the topics within this thesis including spatial audio, binaural audio simulation, auditory motion, perceptual audio evaluation and mixing for picture.

Chapter 3 | Lateral Listener Movement on the Horizontal Plane: Sensing Motion through Binaural Simulation is the first experiment in a three-part study on sensing motion through binaural simulation as compared to real auditory motion in real and virtual environments. An auditory motion recording apparatus was constructed for use in Part 1 and Part 2 of the experiment, as were software evaluation tools. An analysis of the sensation of motion for binaural simulation and actual binaural recordings is given.

Chapter 4 | Lateral Listener Movement on the Horizontal Plane: Part 2 Sensing Motion through Binaural Simulation in a Reverberant Environment is the second experiment in the three-part study. The experiment in this chapter focuses on the influence of reverberance on the sensation of motion, by replicating Part 1's methodology in a reverberant space. An analysis is presented on the sensation of motion for binaural simulation and actual binaural recordings as presented in a reverberant environment.

Chapter 5 | The Effect of Virtual Environments on Localization during a 3D Audio Production Task is the final experiment in the three-party study and focuses on the application of the sensation of motion in real and virtual environments. To do so, this experiment records and analyzes the performance of a three-dimensional audio mixing task which presents auditory motion within real and virtual environments. The experiment uses VR and an HMD for the presentation of the virtual environment.

Chapter 6 | Conclusions summarizes the general conclusions from each experiment along with a general discussion on results and methodologies, and suggestions for future work.

Appendix A provides additional details from the experiment in Chapter 3.

Appendix B provides additional details from the experiment in Chapter 4.

2 BACKGROUND

This thesis includes concepts, methods and experiments inspired by previous research based on the ability to localize sound through binaural simulation and detect auditory motion in real and *virtual auditory environments* (VAEs). This chapter provides an overview of the key principles from prior research as referenced within this thesis. For the sake of brevity, certain fundamental concepts on sound and psychoacoustics including the basic anatomy of the ear, physics of sound, and certain aspects of perception are assumed to be previously known to the reader and/or fall outside the scope of this thesis. For further reading in these areas, the author suggests the most recent edition of Fastl and Zwicker's *Psycho-Acoustics* [4], which contains an extensive overview of fundamental psychoacoustics.

2.1 Spatial Hearing

The topic of spatial hearing is essential to discussing studies on binaural audio, localization, virtual auditory environments and auditory motion. Spatial hearing has been studied at length for over a century [5] spanning foundational works such as Strutt and Rayleigh's *The Theory of Sound* [6] in 1877 (the same year Thomas Edison's Phonograph was invented) to an immense contribution to psychoacoustic research by Blauert, beginning with his Ph.D. dissertation [7] in 1969 to the aptly titled master reference, *Spatial Hearing* [8], and more.

2.1.1 Physical vs. Psychoacoustic

Sound is a physical event, in which a mechanically vibrating source causes particle displacement through pressure waves which radiate away from the source. As with all physical waves, the path sound takes to the a given position (a listener in this case) may be direct or reflected by other objects in space. The sound source (object), its location, and the soundwave it produces, are all physical measures. The ability to *perceive* an object by its sound is a *psychoacoustic* measure. The perception of sound is thereby based on psychoacoustic classifications derived through the detection and alteration of *spatial cues*—perceptual features which are derived through changes in frequency, intensity and time for a given sound source. For a new aspect of the sound to be perceived, there must be a *just noticeable difference* (JND), or detectable change between one spatial cue and another for which a comparison can be made [9] (see also [10]). When a JND is present, a psychoacoustic refinement occurs, further improving the accuracy of the psychoacoustic information. This is essential to one's ability to detect changes in sound position or direction, as explained in Section 2.1.2.6.

2.1.1.1 Auditory Scene Analysis

The result of the evaluation of spatial cues leads to the creation of an *auditory event*, which represents certain characteristics of the physical sound and its physical location in space. Similar auditory events are grouped either simultaneously or sequentially (over time) into one descriptive representation of the sound source, creating an *auditory stream*. The perception of multiple auditory streams and their interaction in space builds an *auditory scene* within the brain, similar to

a visual scene provided by visual cues. One's ability to perceive and group the auditory events transpiring over time within this auditory scene is known as *auditory scene analysis,* named by Bregman [11].

2.1.2 Localization

The human ability to sense sound and perceptually decode it into a classification of the event is certainly remarkable, but as Eargle points out in [12], while one ear allows the listener to determine a wide array of sound attributes including pitch, loudness, timbre and others, two ears are required to determine a sound event's direction and position. This ability to determine a sound source's direction and position is called *localization*, and it's performed by the brain through constant analysis of binaural cues (auditory cues presented at both ears) which are discussed in greater detail in subsequent sections.

In relation to one's spatial awareness, a localizing event provides further separation and classification of sound sources within a multi-stream auditory scene. This separation provides positional anchors (references) for stationary sound sources. Localization also provides the opportunity for comparison between a sound object's previous and current position, which can allude to the motion, trajectory and velocity of the sound object as shown in Section 2.3. It also provides acoustic reference for visual objects in multi-modal situations as described in Section 2.5.

2.1.2.1 Spatial Coordinates

Since localization is a positionally oriented measure, it is important to define spatial positioning,

as experienced by a listener at the listening position. Such spatial positioning is described through spatial coordinates represented by the horizontal, frontal and median planes. It is assumed in experimental research that the human head is symmetrical and that both ears are stationary on the head and positioned away from each other at 180º. The three spatial planes are assumed to intersect with each other at an origin which lies halfway between the two ear canals along the interaural axis which is level with the bottom of the eye socket. The *interaural axis* is a straight line which horizontally intersects the ear canals across the head. The *horizontal plane* thus represents the horizontal angles around the head along the interaural axis. The *frontal plane* intersects the ear canals at right angles to the horizontal plane. The *median plane* lies at right angles to the horizontal and frontal planes, dividing the head vertically between the ears. The ultimate position of a sound source is defined using three polar coordinates: *azimuth,* which describes the horizontal angle, *elevation,* which describes the vertical angle, and *distance,* which describes length away from the listener. Figure 1 details this coordinate system based on an original image by Blauert [8].

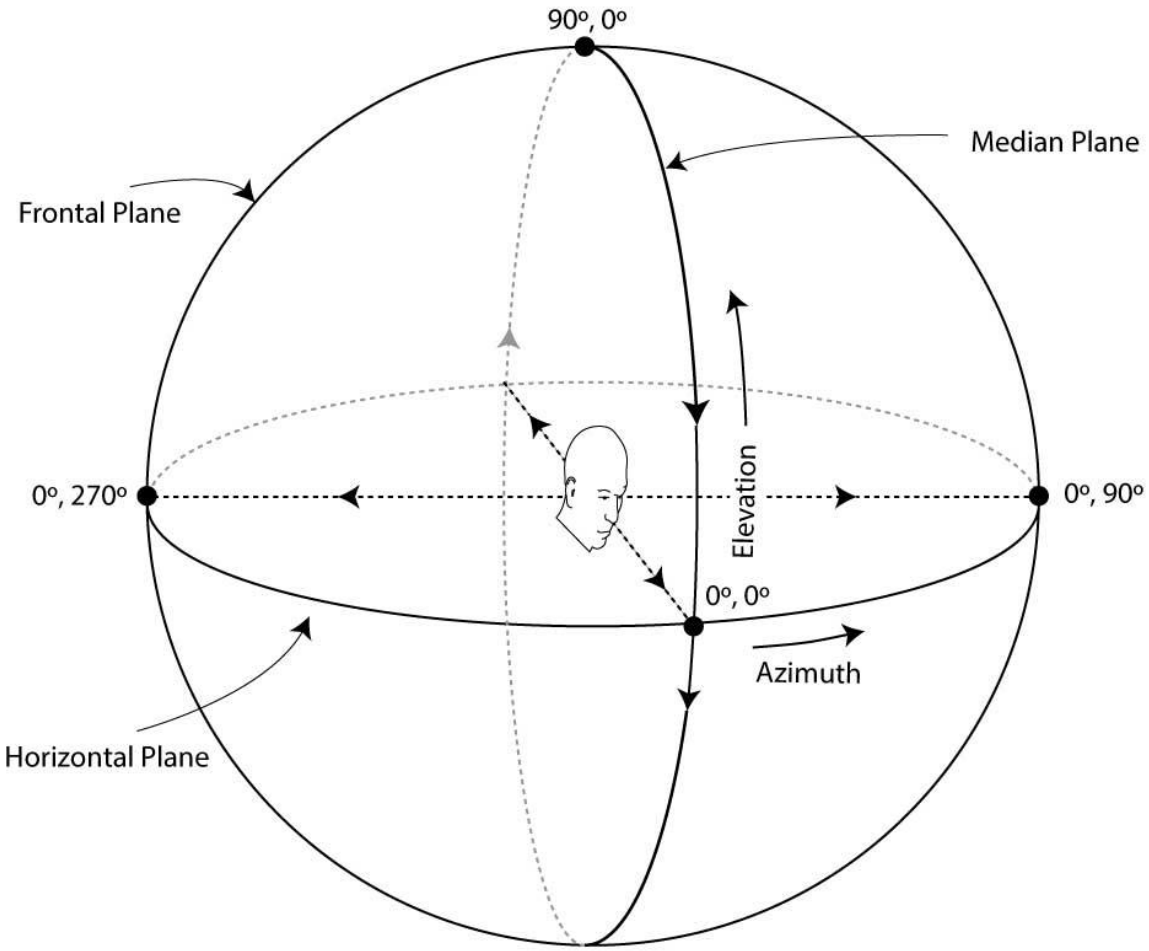

Figure 1: Spherical coordinate system used in auditory experiments

2.1.2.2 Interaural Time Difference

As a sound wave emanates away from the source, it will eventually intersect with the listener. A time of arrival can be measured for the time it takes a wavefront to leave the sound source and reach the listener's ears. Since the ears are separated by a physical distance across the head (along the interaural axis), and since the sound may emanate from an object at some angle away from the

head, the time of arrival for a given wavefront may differ at each ear (Figure 2). Though this difference is incredibly small and ranges in *microseconds* (µs) [9], it represents the most important spatial cue for determining localization on the horizontal plane. This measure is known as the *interaural time difference* (ITD) and provides an ability to localize transient sounds at an angle along the interaural axis relative to the angle of incidence. For a localization angle of 0º or 180º on the horizontal plane (directly in front or behind the listener), the ITD will approach its minimum. A maximal ITD will be observed when the angle of incidence horizontally splits the ear canals across the interaural axis (directly to the side of the head). The average distance between the eardrums ranges from 15 cm to 18 cm [13], which if divided by the speed of sound (343 m/s) yields ITDs of 437 µs to 525 µs for sounds that originate at ±90º on the horizontal plane, however due to shadowing of the head, maximum ITDs can often be slightly larger. While ITDs are highly important to one's ability to localize sound, their effectiveness is greatly decreased for signals above ~1.5 kHz [14].

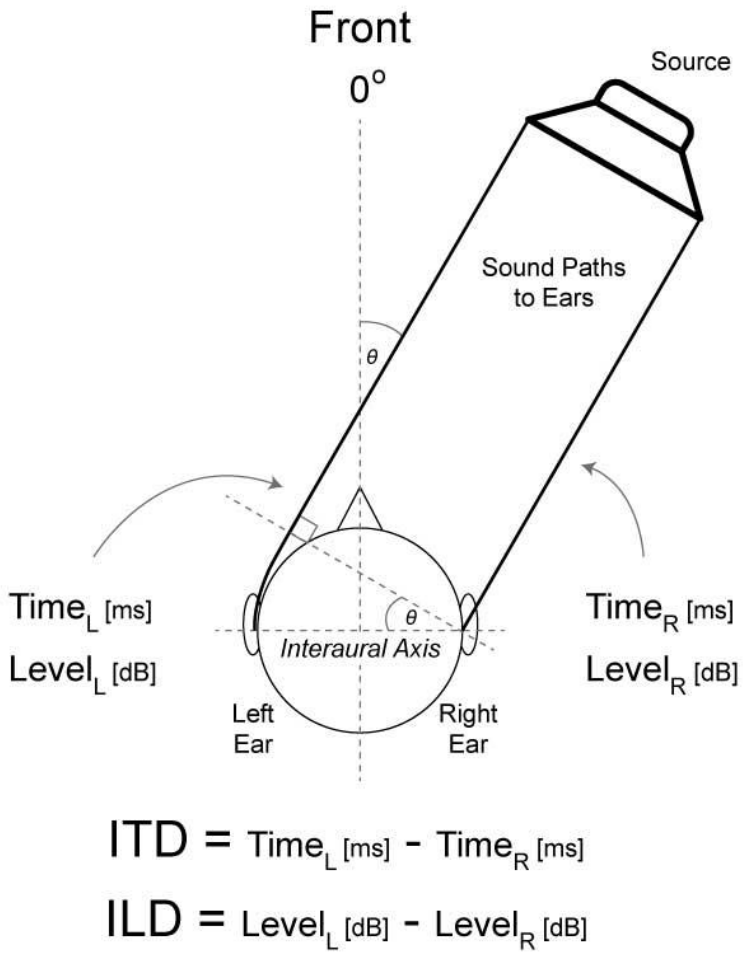

Figure 2: Diagram of binaural cues and the calculation of ITD and ILD

2.1.2.3 Interaural Level Difference

The size of the head and distance between the ears not only presents a sound barrier in which the path of sound is delayed in time between the two ears, but it also creates a difference in the sound's *intensity* (level) at each ear, otherwise known as an *interaural level difference* (ILD). For the sake of visualization, the head can be thought of as a semi-solid circular mass which absorbs and reflects

incoming sound waves. As the sound moves towards the front/back angular positions at ±0º degrees, the ILD at both ears will approach its minimum. When sound waves arrive directly from the side of one ear (±90º), the head will absorb and reflect sound away from the opposite ear and the ILD will be at its maximum. This difference generally occurs at higher frequency ranges where ITD is typically ineffective. The relationship, in which ILD is effective for higher frequencies and ITD is effective for lower frequencies, is called *duplex theory* and was developed by Strutt and Rayleigh in the early 1900s [6]. Maximum differences in ILD occur between 15 *decibels* (dB) to 20 dB.

2.1.2.4 Lateralization

When just one spatial cue is present, rather than an auditory stream containing multiple cues, there is no externalization to the perceived positioning of sound. In this situation, the listener will detect lateral positioning for the sound source, but the sound will not externalize outside of the head. For example, by only receiving an ILD, simple panning will occur within the head from left ear to right and vice versa. However, Jeffress and Taylor found that despite the psychoacoustic differences between *lateralization* (in-head) and *localization* (externalized), the human ability to discern angular direction of a sound source using either cue, results in similar performance [15].

2.1.2.5 Monaural Cues for Height and Distance

While ITD and ILD are important binaural cues for localization on the horizontal plane, other monoaural cues lead to one's ability to localize both vertically and with depth. Early studies by Pratt [16], which were later confirmed by Roffler and Butler in greater detail [17], found a

relationship between audible pitch and vertical position in the *median vertical plane* (MVP) where ITD and ILD theoretically have no influence. Their work defined the *pitch effect*, which refers to a relationship in which humans perceive changes in pitch from low to high frequencies as changes in vertical angle. Pitch Effect is dependent on monoaural cues related to changes in spectral information. For instance, in the MVP, Roffler and Butler showed that a signal with a center frequency of 250 Hz would localize below the median plane. As the center frequency increased, so did localization of the audio image. At a center frequency of 7.2 kHz, localization rose to 20º in elevation [18] [17]. Blauert further observed that these monoaural cues contained *directional bands*, or sharp peaks and dips in the spectrum of a signal which directly related to vertically localized positions on the MVP [7] [19]. Through Blauert's work and others, it has been shown that for sounds with sharp spectral information around 7-9 kHz [14] [20] [21], the sound will be localized above the head.

The perceived distance of a sound object can be discerned by a listener through a perceptual evaluation of the onset and prolonged comparison of direct to reverberant sound, otherwise known as the *direct-to-reverberant ratio* (DRR) [22]. As the direct sound path reaches the listener, it will be immediately followed by the first and early reflection, and finally the reverberant (late) reflections. When the DRR has more direct energy, the sound is closer to the listener. As the ratio progresses to have more reverberant energy, the sound appears farther away. Changes in direct volume also indicate a change in distance. As the sound gradually reduces volume, the listener perceives the sound to move away, and when the sound gradually increases, the sound object is

perceived to move closer to the listener.

2.1.2.6 Minimum Audible Angle

Once a sound source is localized to a position, humans can detect changes in the source's position with a fine degree of resolution through a comparison of spatial cues. This resolution is measured by the *minimum audible angle* (MAA) which is defined as the ability to detect a JND in the perceived change in angle of a sound source. The MAA is a function of the sound source. In the forward horizontal plane where localization accuracy is best, humans can localize sound to an MAA of approximately ± 1º [23] [24].

2.1.2.6.1 Minimum Audible Movement Angle

As it pertains to auditory motion detection in Section 2.3, the *minimum audible movement angle* (MAMA) can be measured as the minimum distance that a sound source needs to move to be distinguished from being stationary [25]. Multiple studies have shown that the MAMA for horizontal, diagonal and vertical movement is 2 to 3 times bigger than MAA's of static sound sources measured under the same conditions [26] [27]. MAMA is a function of velocity and thus is measured in degrees per second (º/s) [28].

2.1.2.7 Localization Blur & Cone of Confusion

Accuracy of localization is measured by *localization blur,* which defines the difference in angular position of the perceived sound to the physical sound source. Figure 3 is based on a diagram of the reported findings in [8] for human localization blur. Localization blur in the horizontal plane is the largest at the sides (90º & 270º) with blur decreasing as sounds approach the front (0º) and back

(180º) of the head. In the vertical plane, localization blur greatly increases as a function of height (Figure 3).

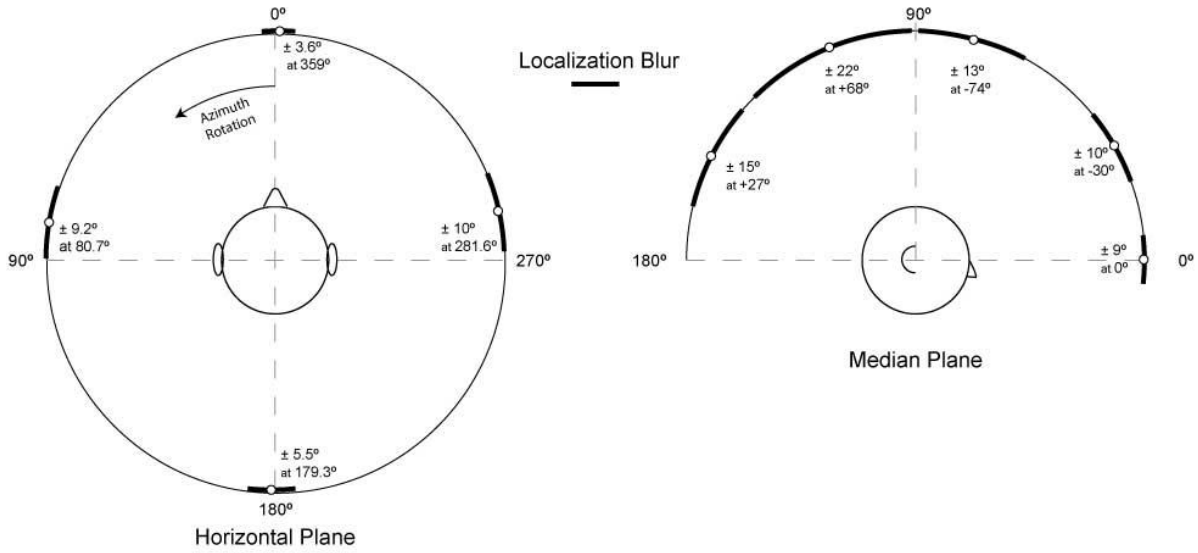

Figure 3: Localization blur, based on a diagram from [8]

As in MAA, localization blur is dependent on the sound source. For single tone or narrow band sound sources, localization blur can be so large that a sound source presented at a side position in the frontal plane, may appear to come from the opposite angular position to the back, creating a perceptual mirror image. This confusion can also occur for sound sources localizing from front to front-overhead. This phenomenon is called *font/back confusion*, which occurs in a hyperbolic area in the horizontal plane known as the *cone of confusion* [8]. To visualize this extreme localization blur caused by the cone of confusion, consider a sound presented at a distance (x) and angle ($ø$) away from the head. Now consider its equal and opposite position on the horizontal plane (x, $-ø$).

As a function of distance and angle, if a sound is presented at either position, the listener will receive the same exact ITD and ILD cues from both positions. Furthermore, in anechoic environments where direct sound dominates and reflections are non-existent, front/back confusion can be at its strongest. Without additional cues such as reflections which add to changes in ITD and ILD overtime, it can be nearly impossible to distinguish which angle of localization is accurate. Only by turning the head can a listener interpolate and compare spatial cues to resolve the accurate localization. Visuals cues related to the sound source can often increase localization accuracy as well, and so it follows that individuals with visual impairments often are better at accurately localizing sound given their heightened sense of hearing.

2.1.2.8 Localization of Multiple Sources

When two or more sound events are presented simultaneously to the listener through multiple sounds sources (i.e., a stereo monitoring system), small inter-channel differences in phase (caused by delay) and level create a perceptual weighting of the signal from one source to the other. This phenomenon creates *phantom images*—auditory fusion of the sound which localizes at some position away from the sources. When inter-channel phase and level values are the same for both sources (the two signals are completely coherent), the phantom image localizes directly to the center position between the two sources. However, as inter-channel differences begin to occur, the phantom image shifts in localization to either source based on the relationship of the differences.

Consider a situation where a listener is sitting in the middle of two loudspeakers. Both loudspeakers reproduce the same signal at exactly the same intensity and phase. When an inter-

channel level difference exists, the sound will localize towards the speaker with the greater intensity. An application of this effect is a stereo pan-pot which positions the sound towards one output by reducing the level of the other. It's easy to see that inter-channel level differences of multiple sources therefore replicate the function of ILD binaural cues for a single source. Conversely, when a slight time delay is introduced at one loudspeaker, localization of the sound event will tend towards the other, less-delayed speaker. This process essentially replicates the time-of-arrival effect of an ITD binaural cue for the listener derived by the inter-channel phase difference.

This phenomenon is known as the *precedence effect* and was first reported by Wallach et. al in 1949 [29]. Wallach showed that in the aforementioned symmetrical stereo listening conditions, the listener will perceive one fused phantom image in the direction of the preceding signal when a signal is delayed at the other loudspeaker by up to 35 *milliseconds* (ms). Beyond this point, the delayed signal is heard as two separate events—a signal and echo. Several others including Lochner and Burger [30] advanced the work of Wallach, however, the most famous work comes from Helmut Haas whose Ph.D. dissertation focused on the intensity of the echo and the listener's perception of these separate events. Haas found "that for [inter-channel] differences from 5 ms to 30 ms, the intensity of the echo loudspeaker must be ten times greater than that of the primary loudspeaker, i.e., against 10 dB, in order to create the impression of equal loudness [31]." Within this range, as level is increased in the echo loudspeaker, stereo widening occurs while loudness remains the same and the echo remains suppressed as a separate event. Compensation of the

localization can also come from increasing the level of the echo loudspeaker. Altogether, this phenomenon known as the *Haas Effect* is widely used as a stereophonic technique to provide psychoacoustic stereo widening through loudspeakers.

2.2 Binaural Audio Simulation

The goal of binaural simulation is to replicate the experience as heard through the two ears (see Figure 4). This experience can be simulated through one's own hearing system, or through measurement techniques which generalize the experience for the purpose of research. For either situation, the human hearing system is as unique to us as our fingerprints. This section will serve as an overview of the necessary concepts required for binaural simulation.

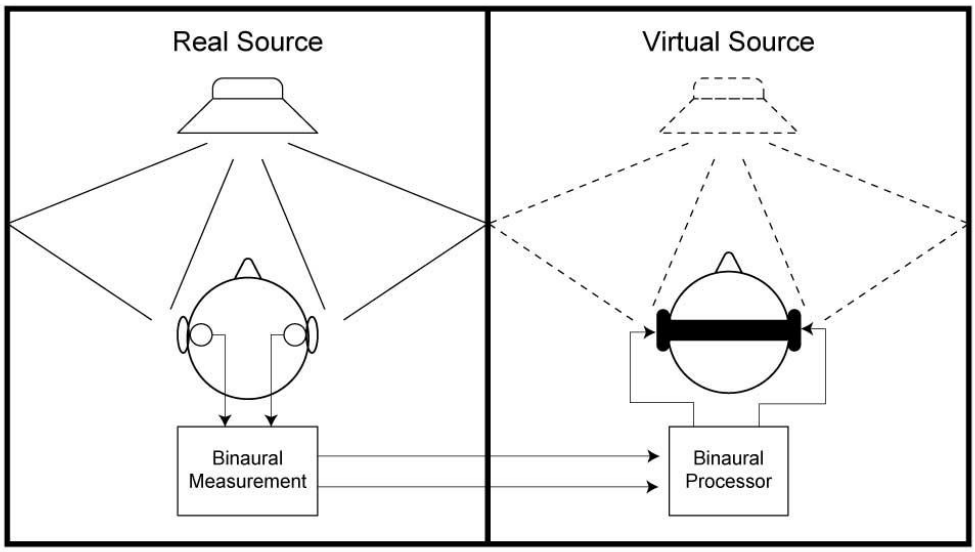

Figure 4: Diagram of the binaural simulation process in which a real source is measured binaurally, then replicated over headphones as a virtual experience

2.2.1 Head-Related Transfer Functions

Head-related transfer functions (HRTFs) represent the unique filtering caused by the head, pinnae and torso on a sound before it reaches the eardrum of a listener in a free field [32]. This filtering effect is due to the reflection, diffraction and absorption characteristics of individual pinnae, as well as the head and torso. It is important to note that HRTFs represent these filtering effects up to a defined position in the ear canal, and in certain applications, HRTFs can include the effects of the ear canal resonance as well. Therefore, it is suggested to define the representative measurement position for an HRTF [33]. The equivalent of the HRTF in the time-domain is the *head-related impulse response* (HRIR) representing the ITD, ILD and spectral cues for a given sound event heard by the listener. HRTFs and HRIRs represent the human hearing system in anechoic listening conditions and exclude the reflective and reverberant characteristics of the room or environment. Figure 5 illustrates an example of an HRTF (frequency domain) and HRIR (time domain) for the same measured location. Looking at the graph of the HRTF, the frequency response for the right ear is largely reduced in magnitude as compared to the left ear where sound arrived first. This is due to the shadowing effect of the head causing a low-pass filter on the high frequency response. The time delay is also present in the graph of the HRIR, where one can see the sound arriving at the right ear is delayed by roughly 350 µs after the sound reaches the left ear. It is also apparent that the overall intensity of the sound event at the right ear is much lower than at the left ear. For perceptual studies in binaural audio, the visual evaluation of HRTFs & HRIRs is an invaluable tool for the visual and statistical analysis of the ITD, ILD and spectral cues that make up a given localized position. Through the HRTF graph in Figure 5, it can also be seen that the magnitude

exceeds 0 dB. This is not to be misinterpreted as distortion. As Hammershøi explains, these overages are mostly caused by “pressure build up and diffraction around the ear, head and body. [34]”

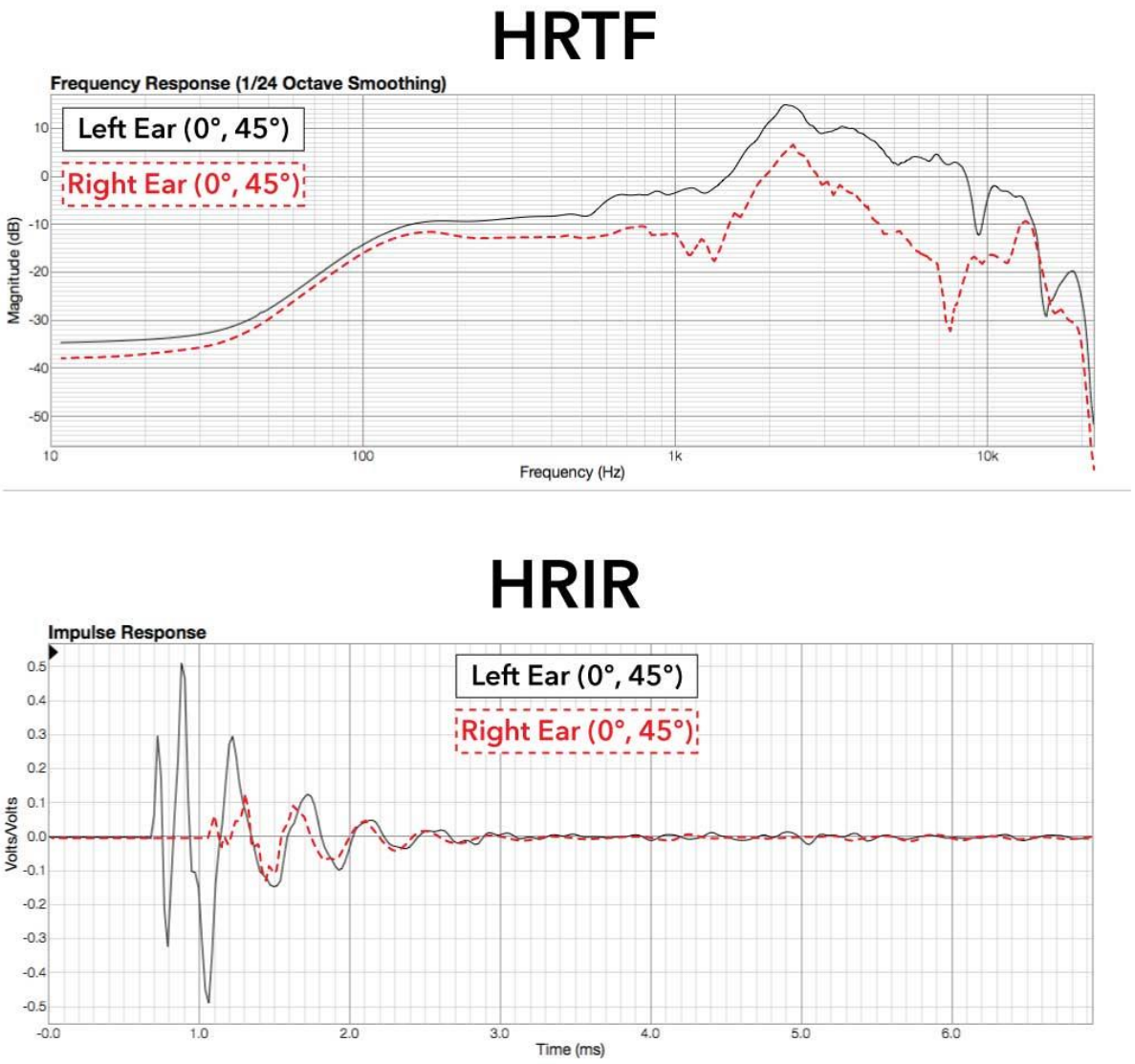

Figure 5: Example HRTF & HRIR for 0º elevation, 45º azimuth

2.2.1.1 Individualized HRTFs

HRTF individualization is a feature of binaural simulation that either needs to be removed or

retained during binaural research, depending on the approach. For example, to understand individualization factors of measured HRTF positions, Hammershøi and Møller studied the standard deviation in subjects for HRTFs measured at three separate ear canal positions [35]:

1. Transmission in the free-field to the block entrance of the ear canal
2. Impedance conversion related to ear canal block
3. Transmission along the ear canal

It was found that all three positions are highly important to individualization, but the lowest deviation values were found at the first, blocked entrance position of the ear canal. They concluded that the blocked-canal position is most suited for HRTF measurements and binaural recordings since sound at this position included the complete spatial information with minimal individualization effects and "may offer a wider range of general applicability [34]."

2.2.2 Measuring HRTFs

The capture and calculation of HRTFs requires extreme precision as to not bias perceptual studies which are dependent on binaural simulation and the ability to provide repeatable binaural cues for localization. In an effort to standardize such measurements, numerous HRTF databases have been created and made accessible for binaural research. These databases contain groups of HRTFs which represent an array of measured angular positions around the head. For instance, the well-known MIT database includes 710 measured loudspeaker positions as individual HRIRs, as well as additional impulses of the loudspeaker in the free-field for measurement equalization [36]. The HRTFs in these databases are generally measured using human listeners, dummy-head

microphones, head and torso simulators, or generated through computer model simulations. The most commonly accessed HRTF databases include the aforementioned MIT database, the CIPIC human-head database of 45 publicly available individuals [37], and the LISTEN (IRCAM) human-head database of 50 individuals [38].

2.2.2.1 Binaural Processing and Convolution

Once the HRTFs are measured, the next step is to process an input signal through the HRTF to achieve binaural simulation. There are two methods to achieve this process. As shown previously, an HRTF provides a frequency domain filter on the input signal, and as such, filter models can be created to replicate the HRTF spectral curve. However, using this method requires additional group delay processing to provide the ITD cues in the binaural simulation. Still, this process is highly efficient as it employs the use of simplified filters and delay lines to achieve the desired spectral and time-domain response. For higher resolution and quality of signal processing of the HRTF, it is recommended to use HRIR convolution to provide binaural processing on the input signal. Figure 6 shows a visualization of the convolution process on a simple input signal. Though the greater mathematical definition and background of convolution is suggested to the reader through [39], the basics of signal convolution for the purpose of this thesis can be summed up through the equation shown in Figure 6, where at a given point in time (t), an input signal $x(t)$ is multiplied by the impulse response of a system $h(t)$ to obtain a convolved output signal $y(t)$.

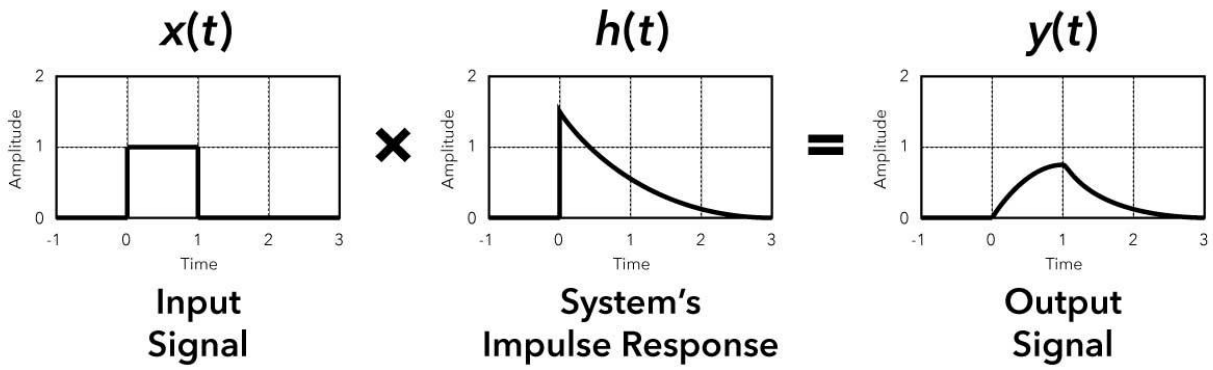

Figure 6: Visual representation of continuous convolution. A square pulse is convolved with the system's impulse response to yield the convolved output signal

2.2.2.2 Impulse Response Capture Methods

A common method for calculating HRTFs is through static HRIR measurements in which small microphones are placed at the entrance or inside the ear canal [8]. Several methods exist to obtain HRIRs. As stated previously, HRIRs contain the acoustic information and resultant binaural cues that are necessary to replicate the linear and time-invariant system between the sound source and the listener's ear. To summarize Møller [40], it is essential to measure signals at maximum amplitude to achieve maximum *signal-to-noise-ratio* (SNR), while also reducing distortion. The method of measurement must result in the highest resolution impulse so as not to distort the binaural simulation, as demonstrated in Figure 5.

Numerous methods exist for capturing acoustical systems through impulse response. The most commonly used techniques are *impulsive noise bursts* (i.e., balloon pop or starter pistol), *maximum*

length sequences (MLS), and the *exponentially-swept sine*. The latter two techniques are the most popular for HRIR measurements as they provide a known input signal to the system and measure the output of the system's response [33] [41] [42] [43]. MLS is a slightly older technique and theoretically provides maximum energy across all frequencies. It is a technique based on pseudo-random noise stimulus. It provides time gating for the measured impulse such that late reflections or noise can be rejected and therefore, gated out. MLS uses a stimulus with more evenly distributed energy giving it higher SNR and reducing problems caused by the random stimulus method [44]. Golay codes, or complementary codes as Cheng reports [45], have been used in an attempt to maintain good SNR while improving the low frequency problems existing in MLS. While it can be a very good method for high quality impulse response measurements, it is highly dependent on system synchronization and high quality digital clocks, since MLS assumes the entire acoustic system is linear and time-invariant [43].

In an attempt to improve upon the MLS technique, Farina developed a method which can simultaneously deconvolve the system's linear impulse response while separating harmonic distortion and noise caused by non-linearity and time-invariance. The exponentially-swept sine technique generates a "sine signal with exponentially varied frequency" and deconvolves the system's response through a linear convolution with an inverse filter of the input [43]. With noise and harmonic distortion separated from the input, "a deconvolution of the system's impulse response [$h(t)$] can then be obtained [by] simply convolving the measured output signal $y(t)$ with the inverse filter $f(t)$", through Equation 1 [43]:

Equation 1: Formula for impulse response deconvolution

$$h(t) = y(t) \ * \ f(t)$$

In general, the key defining elements typically occur at the transient events in an HRIR. These transients occur generally in length of 1-3 ms with the overall HRIR length less than 10 ms [33] [34]. Anything beyond this point would be considered outside the direct response of the human hearing system and towards attributes related to room and environment.

2.2.2.3 Binaural Room Impulse Responses

When the reflective and reverberant characteristics of a room/environment are included in the binaural impulse response measurement process, these measurements are referred to as *binaural room impulse responses* (BRIRs). BRIRs are often used to study the influence of room acoustics on localization [46]. BRIRs are also used to create more realistic binaural simulations over simulated reverberation, or to predict real-world performance of a binaural simulation [47] [48]. Figure 7 illustrates the time domain comparison of an HRIR pair (without reflections) and the BRIR pair with noticeable reflections caused by the environmental acoustics.

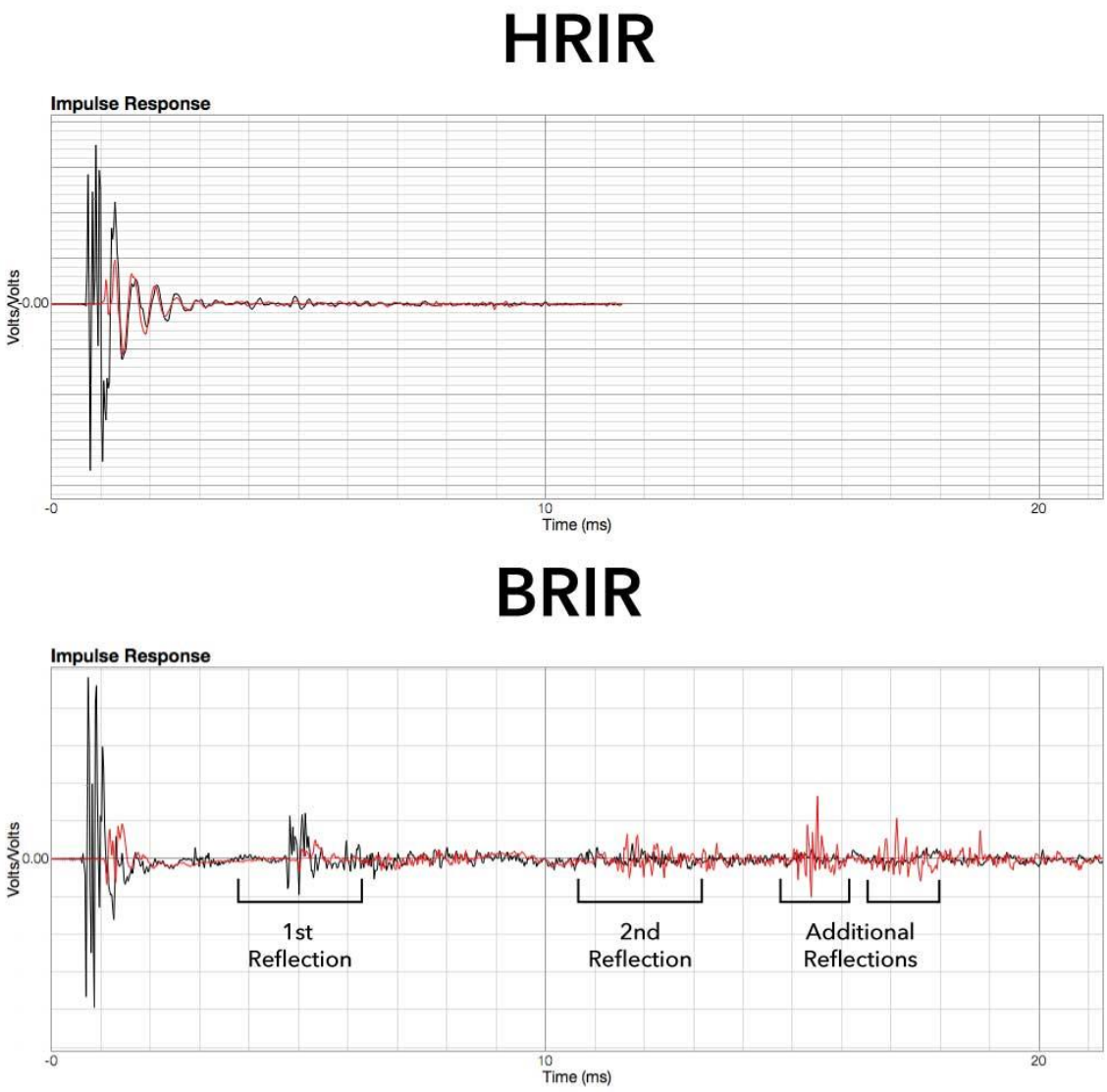

Figure 7: Diagram of HRIR vs. BRIR pairs (black = left ear, red = right ear) within the same time window, measured at 0º elevation, 135º azimuth. HRIRs exhibit no significant reflections after the first transient event. BRIRs exhibit numerous reflections caused by the room

2.3 Auditory Motion

Much like localization, the auditory motion percept is vital to our survival. Motion can be observed through multiple sensory modalities, but historically, most research has focused on visual information [28]. In fact, studies on visual motion dominate the topic of research in terms of available publications such that there is substantially more work focusing on *apparent visual motion* (optical illusions) than auditory motion [49]. However, due to the recent introduction of high quality virtual and augmented reality technologies, the field of auditory motion has seen a necessary increase in published research. This section will focus primarily on the auditory sense as it relates to object motion and the illusion of self-motion. Typical methods used to simulate auditory motion through auditory cues will also be discussed.

Auditory motion is the study of movement perceived by a listener through auditory cues. The auditory motion cues represent changes in position of the sound source (i.e., through different localization cues), and are detected through a fundamental comparison process. To summarize Carlile [28], there are three perceptual attributes to a moving sound source that must be considered:

1. Is the source moving or stationary?
2. What is the location and trajectory of the motion?
3. What is the velocity of motion relative to the head?

Though auditory motion is heavily dependent on spatial localization, interestingly, auditory motion

cues are not always limited to the binaural process. Strybel & Neale have found that monaural processes can also present auditory motion through spectral changes relating to different locations [50]. Further to this point, Phillips and Hall [51] found that the perception of motion is dependent on temporal windows, in particular, a *stimulus onset asynchrony* (SOA) between 30 – 60 ms, but while motion can be perceived monaurally, binaural cues are likely required to determine the direction of motion.

2.3.1 Auditory Apparent Motion

Apparent motion describes the illusion of motion that occurs when stationary objects, presented in rapid succession or in relation to other moving objects appear to move. A great example of apparent motion is an animated film where static images are presented in succession to give the illusion of movement. In the auditory context, it follows that *auditory apparent motion* describes illusory motion induced by the succession of auditory cues. Like actual auditory motion, Strybel has reported that auditory apparent motion can also be induced through monoaural and binaural cues [49] and that this phenomenon seems to occur at *interstimulus onset intervals* (ISOIs) of 20 – 130 ms, while Briggs and Perott detected motion through ISOIs as low as 10 ms [52]. Both reports are consistent with the results found in some of the earliest experiments on auditory apparent motion by Burtt [53] and Klemm [54], which are referenced in a large methodological review conducted by Mathiesen in the early 1930s [55]. Mathiesen developed a methodology that found results which somewhat contradicted the overall plurality of auditory apparent motion that was found by Burtt in his subjects, though she too, found evidence of apparent auditory motion.

For example, Burtt found most (4 out of 5) observers experienced motion phenomena in his experiments, whereas Mathiesen had very few observers (mostly observers “D” and “I”) state that they experienced movement across her initial experiments. Further to this point, Mathiesen never explicitly asked the observers to report the motion they perceived in the first set of her experiments, which was done in an attempt to replicate the methodologies of precursory studies. Instead, her instruction to the observers was this:

"A few seconds after the signal an auditory stimulus will be presented, your perception of which you will be asked to describe."

It was through the analysis of the observers' comments in how they described their experience (in addition to their results) that she was able to confirm that some observers had experienced a motion phenomenon. Much of the motion phenomena was related to what Mathiesen describes as "fillings" (sound that appeared to be different or a continuation of the two switched sounds presented between two receivers). It was only until Mathiesen offered the suggestion of motion in her instructions for later experiments that observers reported motion more frequently.

Mathiesen's work, along with the other early studies performed by those she references such as Burtt’s, had evaluated auditory motion through perceptual "switching" methodologies—turning one source on then switching off to play another. These studies did not present listeners with crossfading techniques which offer continuously mixed ramps of level over time similar to the panning techniques used in audio productions. Therefore, it should be noted that Mathiesen also

reported certain perceptual conditions due to switching, such as long interval length, which often presented "two-ness" (meaning two distinct sound sources were heard)—a phenomenon caused by this perceptual switching. Mathiesen's study is a great foundational work for the consideration of experimental design and development of best practices in auditory apparent motion experiments. It is suggested for further review by anyone considering research in the topic.

As discussed later in Section 2.5.2, temporal asynchrony of auditory cues along with visuals can also significantly affect visual illusions of apparent motion. Consequently, it is important to understand in the context of audio-visual applications, that auditory cues may lead to unwanted motion effects if paired improperly or may enhance the perception of motion if properly aligned with visual objects.

2.3.1.1 Auditorily-Induced Vection

For the purpose of audio production, it is important to understand to what extent auditory motion cues facilitate apparent motion, and specifically, *apparent self-motion* (vection) as it directly relates to the success of first-person virtual reality applications. A classic example of apparent motion describes a person on a parked train in the station feeling as if they are moving backward when the train on the neighboring track beings to move forward. As Boring noted, such illusions of motion, like in this example, can produce a real "*sensation of movement*" [56]. However, self-motion is much more effective in the visual modality than through audition. Urbantschitsch stated in one of the earliest known reports on auditorily-induced vection [57], that the "self-motion sensation is so slight that one should pay special attention to notice it." To this exact point, Riecke

conducted multiple studies on the topic to understand how auditory cues may facilitate vection. In one such study, using monoaural and binaural auditory cues, he paired the auditory cues to matching visuals within a rotating virtual reality environment (known to provide vection). The results showed that the binaural auditory cues significantly enhanced vection and presence whereas monoaural cues did not significantly affect any measures of vection [58]. In a similar study looking at linear vection through translational object motion, Calabro et al. confirmed Riecke's findings that binaural (spatialized) auditory cues enhanced vection when coupled with the visual cues in the visual scene [59]. These findings suggest that however slight the self-motion sensation may be, it still is significant in providing assistance to the overall perception of self-motion.

2.3.1.2 Binaural Motion and the Sensation of Motion

Chapter 3 and Chapter 4 of this thesis describe a novel methodology for translational auditory motion through binaural simulation. For clarity over the multiple classifications of auditory motion, the term *binaural motion* is used by the author to describe the process for simulating and recording auditory self-motion through binaural processes. Further to this point, the author uses the term *sensation of motion* as a metric which represents the degree at which motion is perceived by a listener through the presentation of binaural motion.

2.3.2 Auditory Motion Methodologies

To study auditory motion, most researchers simulate or capture auditory cues through one of two experimental methodologies: Rotational Sound Sources and Listener Rotation (see Figure 8). In both setups, all sound sources are typically equidistant from the listener to remove changes in

loudness as a function of the results.

2.3.2.1 Source Motion through Rotational Sound Sources

This method rotates one or more acoustic sound source around a static listener seated at the center of a spherical speaker array. The speakers and listener remain in a static position as sound sources are panned through the speaker array to present a simulation of auditory motion. In this scenario, only the sound source moves. This method is preferred for most auditory motion and localization studies as it ensures the listener position is absolutely still and unchanged. It also ensures the sound source follows a controlled, predictable path.

While the entire physical system is stable, this method relies on continual source distribution and proper phantom imaging between physical loudspeakers to achieve a full range of simulated source motion. *Gain factoring* (distribution of the acoustic source energy) must be accurately calculated when a source is panned from one loudspeaker to another. For this reason, several methods and algorithms for acoustic source panning have been designed in an attempt to simulate real-world acoustic source movement [60] [61] [62]. Some of these methods are covered later in this section. Recent studies by Frissen [63] and Sankaran [64] have used this method to study the detection of motion at varying velocities by rotating sources at increasing and decreasing velocities around a listener.

2.3.2.2 Physical Motion Through Listener Rotation

The second method rotates the listener on a central pivot point (using a circular platform) within the spherical speaker array. This method eliminates any potential flaws in acoustic source

distribution mentioned in the rotational source method. The rotating listener is provided auditory cues through direct sources only. Source interpolation is purely the result of auditory cues caused by the change in the listener's position, relative to the direct sound source. This method has been used recently by Riecke [65] & Nykänen [66] to study the topic of self-motion in complex VAEs.

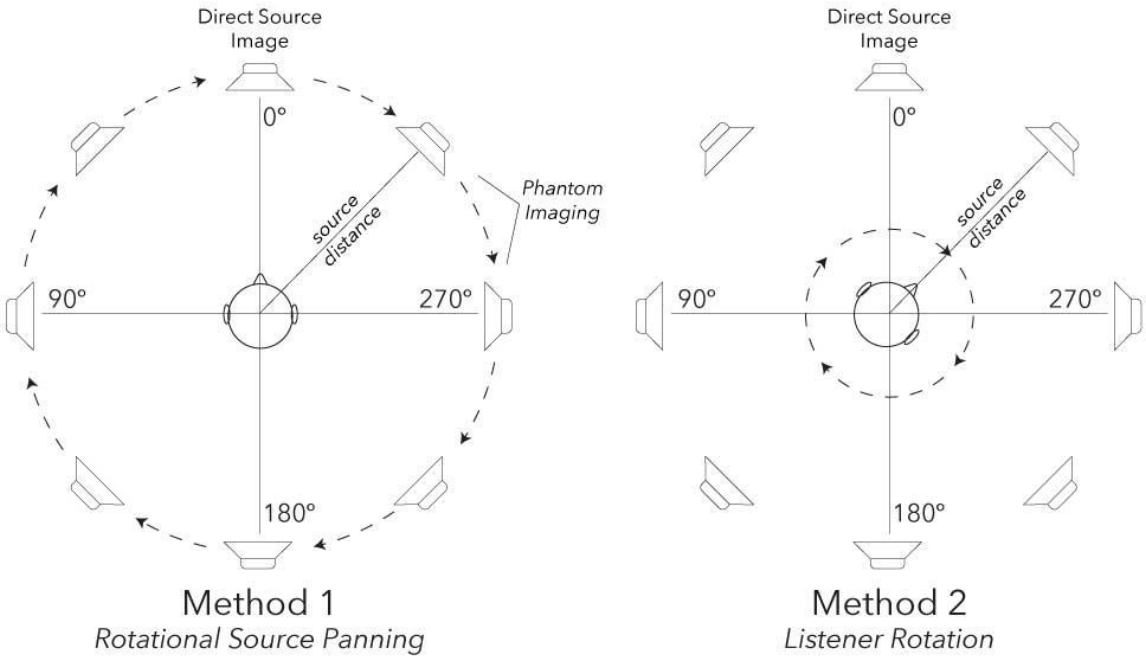

Figure 8: Methods of source and listener rotation for auditory motion experiments

2.3.3 Perception of Relative Movement

The perception of motion is a complex and complicated topic to study since it requires a continual comparison of positional cues relative to one's own physical position and movement in space. To a rotating listener in empty space, the forward perspective equates to all events occurring directly

in front of one's face—in the direction of 0º azimuth, perpendicular to the interaural axis (see Figure 1). Obviously, the world is not full of empty space but instead filled with static and moving objects. So, it's interesting that despite this complexity, from our own perspective our auditory world appears to be stable, despite the fact that at any given moment we are evaluating our world through continuous head movements.

Motion perception presents an interesting challenge to researchers since it is both rotational (angular) and translational (longitudinal or latitudinal). Further to the point, relative position to static objects must be considered during any motion simulation or study as a simple movement in the listener can cause a change in the angular calculation of an object's motion. For example, when head rotation occurs, everything surrounding the listener must be re-calculated based on relative angles and distance. Figure 9 demonstrates this complexity in the auditory sense whereby a car is driving along an arc path while a listener starts at some distance away from the car then walks while rotating to meet the car at its final destination. Though many variables are left out in this diagram for simplification, it is easy to see the mathematical complexity that exists in this auditory motion scenario for anyone attempting to simulate this real-life situation in virtual reality.

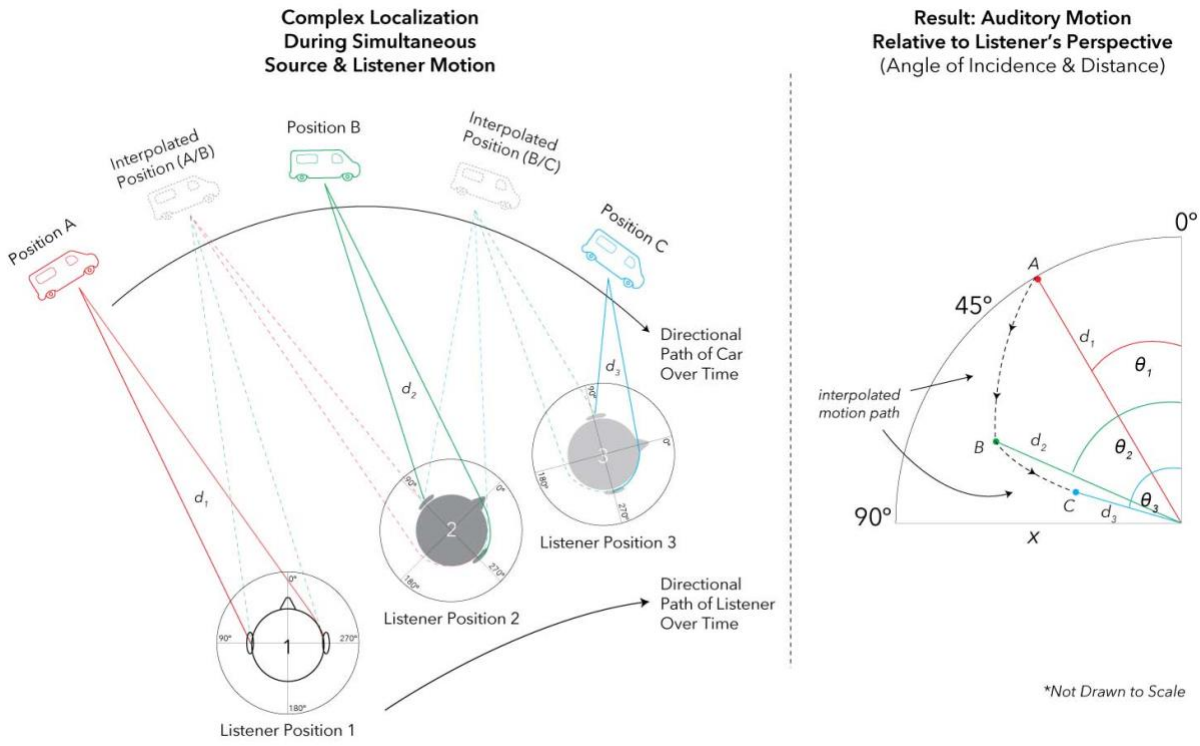

Figure 9: The first diagram shows complex localization during simultaneous source and listener movements. A car at position A is heard by a listener at position 1. When the listener reaches position 2, the car is at position B and so on. The second diagram shows how auditory self-motion can be simulated by interpolating the measured spatial cues at each listener position where d_1, d_2, and d_3 are the distances (yielding time of arrival differences) from listener position to respective source position, and θ_1, θ_2, and θ_3 are the relative angles of incidence at each listener position to respective source position

2.3.4 Techniques for Binaural Motion

Simulating real-world auditory motion requires specific techniques to replicate the continuous

change in the wavefront's angle of incidence, as well as its respective distance from the listener. This process typically requires interpolating positions to allow for the sound source to perceptually "move" from one source to another. As Figure 9 shows, source interpolation must occur if one hopes to simulate the auditory cues between Position A and B, or Position B and C. For example, as the listener observes the car moving from A to B, the perception of this inter-positional movement is equivalent to the simultaneous playback of two virtual sources, with source (A) steadily decreasing in volume (after the car leaves position A), and a source B steadily increasing in volume (until the point the car arrives at position B). For anyone experienced in stereo reproduction, this process is akin to twisting a pan pot to left or right—the operator essentially reduces the output to the left channel, or to the right.

Another point to consider in a moving source/listener scenario is the difference in distance that exists between the source and the listener at one position to the next. Distance directly relates to the time of arrival, and a difference in time of arrival is perceived as translational motion (forward/back or side/side), in which the sound appears to move toward or away from the listener [8]. In the case of binaural simulation, binaural cues and time of arrival must be interpolated in real-time, which can lead to some potentially complex *digital signal processing* (DSP) challenges. While this area of study is quite vast, the following sections will serve as a general introduction to the basic methods commonly used for motion interpolation of binaural signals.

2.3.4.1 Binaural (Direct HRTF) Switching

In many cases, the comparison of auditory cues that lead to auditory motion can be thought of as

a perceptual "switch" from one localized position to the next. Binaural switching, or direct HRTF switching, does just that; it is a method for motion simulation by processing an input signal through sequentially selected HRTFs. The process is relatively simple and computationally efficient. It can be equated to the *single pole, double throw* (SPDT) switch in an electrical system. In a binaural simulation, an input signal is convolved through an HRTF representing point A, then a direct switch is made, and the signal is redirected and convolved through another HRTF representing point B (Figure 10).

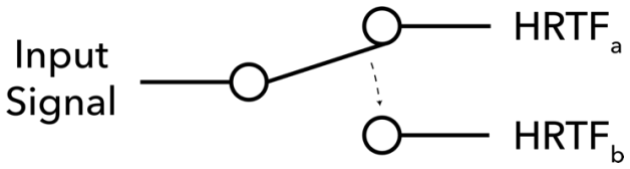

Figure 10: Diagram of a simple binaural switch. $HRTF_a$ represents a position (A), and $HRTF_b$ represents a separate position (B)

However easy it may be, HRTF direct switching naturally presents some problems. First, the process involves changes in temporal (ITD) and spectral characteristics and as such, directly switching between the two convolved signals (i.e, two HRTFs) can therefore lead to time-variant switching artifacts in the form of audible clicks. Hoffmann and Møller investigated this problem by studying the detectability of audible artifacts (clicks) through the *minimum audible time switch* (MATS) which defines the "smallest pure time switching that causes an audible artifact" when making a switch between ITDs related to different HRTFs [67]. As a measure of perceptual resolution, MATS values can inform the successful simulation of motion for which time switching

artifacts will go undetected by the listener. They showed that detection of time switching artifacts is greater in pure tone than broadband noise, which is believed to be due to the masking effect of noise over the artifacts as opposed to clearly heard artifacts added to the pure tone. Overall, the lowest MATS for pure tone signals was 3.6 µs at 90º left on the horizontal plane. Their data also shows that for a broadband signal at all spatial locations, the MATS threshold is 5 µs, which suggests that motion simulation algorithms using direct HRTF switching must stay below this value to go undetected by artifacts.

2.3.4.2 Binaural Interpolation

Resolution is another concern with direct switching because switching from one simulated location to another may not give a sense of fluid movement, but instead an unwanted perceptual "jump" in position. To solve this problem, interpolation methods can be used, especially in a situation where an HRTF database contains less measured HRTFs than the experimental design requires. It is also important to consider the quality of the interpolation method, as well as the detectable resolution (angular representation of measured and interpolated HRTFs). Chapter 3 references a common motion interpolation method for HRTFs called *inter-positional transfer functions* (IPTFs) and can be visualized in Figure 9. Freeland's IPTF method is a scalable real-time process for "small displacements" in which the unknown midpoint ($IPTF_{i,f}$) along the path from a known initial position ($HTRF_i$) to a final position ($HRTF_f$) can be interpolated through a positional ratio of the two transfer functions [68]. While this method is computationally sound and efficient and expands upon the lower resolution bilinear interpolation method [69], Freeman failed to compare this

motion simulation technique to any real-world binaural motion recordings, in order to understand subjective analysis of such interpolation. Consequently, this method has inspired the work of this thesis as a basis for subjective analysis in binaural motion simulation.

When interpolating any HRTF or motion path, the resolution of HRTF measurements in any database must be considered in terms of perceptual resolution requirements. For example, a study by Minnaar, Plogsties and Christensen evaluated the VALDEMAR database of nearly 12,000 measured angular positions represented through HRTFs [70]. The HRTFs were separated by a resolution of 2º. The researchers looked to determine the necessity of all of these HRTFs by evaluating the detectable audible difference between interpolated HRTFs against measured HRTFs of the same location. HRTFs were interpolated over large range of angular spacing between measured HRTFs by linear interpolation of the minimum phase components in the time domain, while retaining the ITD values from the measured HRTFs. Their study found that for static positioning and motion simulations, the necessary resolutions for interpolation in the median plane and frontal plane range from 16º down to 2º, as a function of elevation, with only 24º of resolution necessary for overhead. A resolution to the sides of the head (on the horizontal plane) is suggested at 4º - 8º. Their results could help to "minimize the number of HRTFs that have to be measured" and also "optimize the spacing of HRTFs [when measured] at high resolution." Their work reduced the original 12,000 HRTFs down to just 1,130 spatially necessary HRTFs. A very recent study by Sloma et al. furthered the concept of complex motion simulation through HRTF interpolation while reducing the number of real measurement positions in the interpolation process, and

including the addition of sound source directivity information in the binaural simulations of motion [71]. Sloma concluded that the presence of sound source directivity information in their binaural simulation created a more realistic binaural experience.

2.3.4.3 Binaural Crossfading

A method that is derived from the interpolation of HRTFs is referred to as binaural crossfading. This technique is the same as any crossfaded signal process in which 2 or more signals are combined by proportionally raising or lowering each signal's intensity through *linear or logarithmic time ramps*. Typically, this process is done at the convolution stage, where an input signal is convolved with the HRIRs involved in the interpolation process, and the output intensities (overall gain) of these convolutions are controlled through the crossfading process. As Enzner notes, crossfading is used for "its simplicity in many practical applications" of VAEs [72] (see also [73] [74]). This signal process is illustrated in Figure 11.

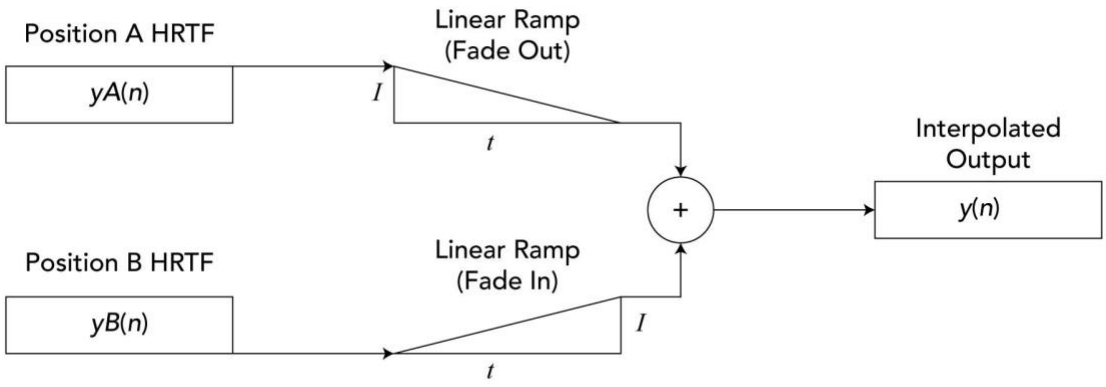

Figure 11: Diagram of simple binaural crossfading technique. *yA*(*n*) and *yB*(*n*) represent the

convolution process for the input signal at position A and position B. Both signals are fed into a linear ramp to proportionally fade each signal's intensity (*I*) over time (*t*). The addition of both signal's intensities results in the combined, interpolated output convolution

Matsumoto and Tohyama suggested the simple linear crossfading technique was "not a realistic way to represent a moving sound image" since it assumed that motion had to be simulated by two simultaneous sound sources [75]. In response to this problem, they expanded upon the IPTF method [68] to create an algorithm which simulates motion through linearly interpolating both, BRIRs and the time of arrival difference between the BRIRs and their respective source measurements. This interpolation technique was subjectively measured against a real-world recording of the same motion which was acquired through a *head and torso simulator* (HATS) by using the listener rotation method shown in Figure 8. During the recording, the HATS rotated while sound was produced from one statically positioned source in an anechoic environment. The results showed that their algorithm for interpolating motion provided such realism that subjects could not discriminate between real and interpolated source images. Matsumoto's methodology proved successful for comparing the perception of real-world binaural motion against a binaural simulation. However, it failed to look at the influence of reverberant environments on the perception of motion, as well as the interpolation of translational motion between two active sound sources. This work inspired the experimental design and methodology in this thesis and was the basis for the hypotheses.

2.3.4.4 Vector Base Amplitude Panning

For higher resolution auditory motion in three dimensions, binaural crossfading may not be sufficient and so, a gain factoring method is used across the entire VAE. A commonly used solution for positioning a sound source is the *vector base amplitude panning* (VBAP) method. In a basic description, the VBAP method processes an input signal with separate gain factors at each individual output relative to an interpolated source, and is calculated as a linear combination of source vectors where the base is defined by unit-length vectors pointed towards the sources from the listening position [61]. To be more specific, the direction of the interpolated sound source is defined by a three-dimensional unit vector representing each unit vector of the 3 separate neighboring sources. The localized source image is thus a factor of all the neighboring gains (this would equate to individual levels of HRTF convolutions in the binaural simulation context).

While VBAP is a great method for spherical (3D) panning, Pulkki notes the disadvantage of "all amplitude panning methods [including VBAP, is that] the virtual [interpolated] source cannot be positioned outside the active arc or region," and only appears at the same distance away from the listener as the measured sources [61]. This makes VBAP a poor choice for translational motion simulation.

2.3.4.5 Doppler Effect

An important process to consider in any motion simulation is the phenomenon known as the *Doppler Effect* which occurs when a stationary listener observes a rise and fall of the perceived pitch for a sound produced by an object that approaches them, then passes by. Anyone familiar

with the sound of an ambulance approaching and passing by has experienced this illusion. As Farnell describes in [76], the doppler illusion occurs because "the speed of sound (c) remains constant but movement [of the sound-emitting object] towards the listener squashes the wavelength of the sound by $(1 - v/c)\ \lambda$, and for a source moving away the wave is stretched by $(1 + v/c)\ \lambda$." Therefore, Farnell suggests Equation 2 can be used to simulate the Doppler Effect, where (+) indicates movement toward the listener and (–) indicates movement away from the listener, at a velocity (V):

Equation 2: Formula for doppler effect simulation

$$f_{\text{observed}} = \frac{c}{c \pm V_{\text{source}}} \times f_{\text{source}}$$

2.4 Perceptual Evaluation for Auditory Motion

Methods for the perceptual evaluation of auditory motion (for the detection of motion) are roughly based on audio quality and localization assessment methodologies. This section illustrates how little attention has been paid to establishing specific methods for comparing auditory motion stimuli during listening tests. While standards still need to be set, widely established methods for assessing audio quality already exist and have served to influence the design of recent auditory motion listening tests [77] [78] [79]. Accordingly, some suggestions have been made within the literature for the design of such auditory motion experiments, including those used within this thesis. For instance, Väljamäe notes that it is common practice in vection research to use subjective scales when accessing the strength of the perceived self-motion [80]—scales related to audio

quality assessments. Furthermore, in a recent streaming VR experiment, Narbutt et al. found compressed spatial audio (of decreased quality) can significantly inhibit localization accuracy—the primary component of auditory motion cues [81]. This section introduces the most commonly used perceptual audio assessment methodologies as they relate to auditory motion assessment and discusses these methods for how they may influence future listening test design in this field.

2.4.1 Considerations for Perceptual Listening Tests in Auditory Motion

As Rumsey points out, "listening tests are the most widely used formal method of sound quality evaluation" [82]. The perceptual evaluation of audio is a highly subjective process, and as such, listening tests require the researcher to be conscientious with regard to their selection of listening subjects. Throughout the literature, numerous studies have concluded that significant differences exist when the evaluation of audio quality is performed by trained groups of listeners vs. untrained listeners [83] [84] [85]. Further to this point, trained or experienced listeners are more reliable in judging audio quality than untrained, novice listeners [86], specifically those experienced in audio production and musical training [87]. However, most, if not all of these studies are based on evaluating *audio quality* rather than *auditory motion detection*. Historically, auditory motion detection and localization listening tests have been performed by generalized groups of listeners spanning from inexperienced levels to experienced. When evaluating the influence of audio and video on the quality of an audiovisual presentation, Kohlrausch found that "audio impairment is more detrimental to quality assessment than a video impairment" [88]. This would suggest a link between quality and motion since the audiovisual presentation contained auditory and visual

motion cues. Therefore, care must be taken when selecting listeners for auditory motion perceptual listening tests, such that well-established methods for the evaluation of audio quality serve as an influence for listener selection and experimental design.

Additional factors should be considered for the selection of listening subjects. For example, although Olive found differences in performance for experts and novices evaluating audio quality, his same study found the all listeners were equally reliable when evaluating a measure of preference in an audio system [84]. Historically, auditory motion experiments have used generalized listener groups, since the detection of auditory motion also relates to localization, as noted in [55]. In this case, it is best to use general listeners because the task is to understand how the general population localizes binaural cues, rather than how an expert can determine the precise audio quality of a localized sound object.

Age is another consideration. Howie makes reference to a study by Fortenbaugh which found that individuals past the age of 43 have decreased sustained attention ability [89]. Howie notes that these findings, along with his own experimental results suggest that an "optimal age range" may exist for listeners in their mid-30's due to one's ability to focus on the sustained task of listening, as well as the fact that "subjects in this age-range should not yet be suffering from age-related hearing loss" [87]. Carlile also points out that while "no previous experiments have directly addressed the impact of hearing loss on spatial motion perception," it has been widely demonstrated [90] that static localization test "performance is generally decreased for the aging population" and for "people with hearing impairment" [28]. This all suggests that motion

evaluation can be evaluated reliably by general listener groups, however care should still be taken to train the listeners of the test stimuli and application, as well as to exclude listeners with hearing loss.

2.4.2 Methods for Detecting Quality Differences in Auditory Stimuli

The following subsections detail the International Telecommunication Union (ITU) standards designed for detecting audio quality differences in certain listening applications. Rumsey points out that "human listeners tend to be quite poor at judging sound quality attributes when they have nothing to compare with, but can be quite reliable when there is one or more 'anchor' points on the scale concerned" [82]. These standards address this point exactly, as both are based on the use of audible anchors and quality-rating scales.

2.4.2.1 Double-Blind Triple Stimulus and Hidden Reference

ITU-R BS.1116-3 introduces a listening test method to assess the *basic audio quality* (BAQ) of audio systems which introduce small impairments to an ordinal audio stimulus [79]. It consists of a double-blind test using three stimuli. A hidden reference and the affected stimulus under test (the condition) are randomly presented to the listener as sound sources "B" and "C." B and C are then compared to "A" (the marked reference) using a *continuous quality scale* (CQS) from grades 1 to 5 to reflect the impairment of the affected audio. This process of subjective comparison asks the listener to determine the audible difference between the reference stimulus and the condition stimulus under test. The hidden reference stimulus compared to the reference should exhibit an *imperceptible* difference or impairment (assigned a grade of 5) where a severely affected condition

will be rated as *very annoying* (assigned a grade of 1). Figure 12 illustrates the CQS recommended for ITU-R BS.1116. This method is recommended for use with listeners who have experience in listening to sound critically and are able to detect the fine perceptual details of impairment within an audio signal.

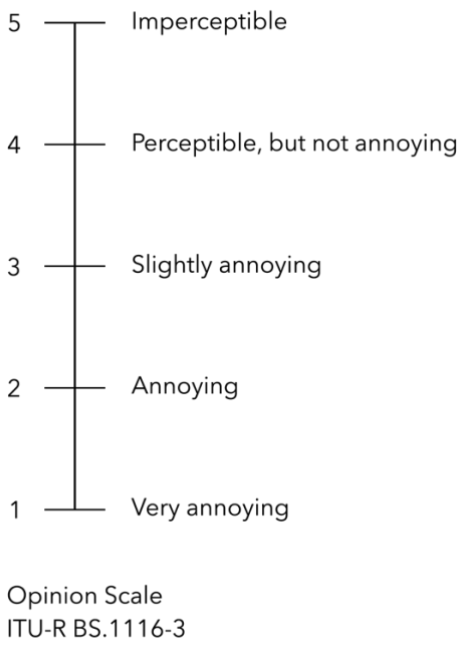

Figure 12: Continuous quality scale recommended for ITU-R BS.1116-3

2.4.2.2 MUSHRA

ITU-R BS.1534-3, also known as MUSHRA, which stands for *Multiple Stimulus with Hidden Reference and Anchor*, is a quality scale evaluation methodology originally developed for the evaluation of perceptual (lossy) codecs such as MP3s, to determine perceived scale of quality for

various bitrates [77]. By design, MUSHRA is a double blind, multiple stimulus test in which listeners are presented multiple conditions for which they must rate against a hidden reference stimulus of the highest quality. Among the stimuli are two hidden anchors of medium and lowest quality. The test allows the listener to switch between conditions in real-time and grade each condition on a CQS. Figure 13 illustrates the CQS recommended for MUSHRA.

With MUSHRA, it's is easy to see how adaptations of this methodology can be applied to the subjective detection of other impairments, such as those in localization or motion stimuli when compared to a reference. For instance, in Narbutt's previously mentioned study, MUSHRA was used to evaluate spatial audio compression as an impairment on quality and localization accuracy [81]. Like ITU-R BS.1116, the ITU recommends that MUSHRA be performed by listeners experienced in the critical evaluation of audio signals.

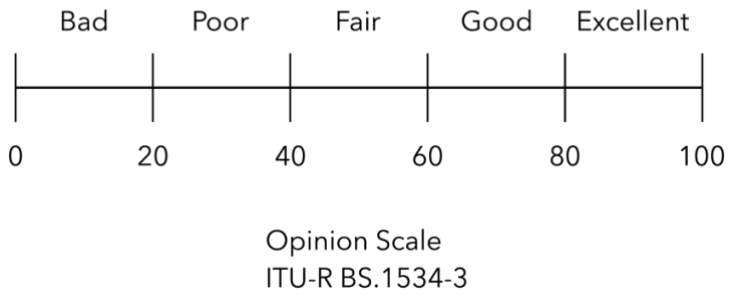

Figure 13: Continuous quality scale recommended for ITU-R BS.1534-3

2.4.2.3 Other Considerations in Detecting Audible Differences

ITU-T P.800 recommends the use of an *absolute category rating* (ACR) scale for when subjective assessment of the difference in audio signal quality is best graded through absolute measures [78]. In this method, the listener grades their opinion of each stimuli on a scale of 1 – 5 where 1= "bad", 2 = "poor", 3 = "fair", 4 = "good" and 5 = "excellent." Results should be reported as the mean opinion score (MOS).

2.4.2.3.1 Sample Familiarization

In addition to training listeners on the process of performing the listening test through a trial experiment, it may be helpful to provide them with an additional awareness of the range of samples within the test. For instance, when aspects of the perceptual listening test require prior training on the features of the stimuli or the group of individual stimuli themselves, it can be beneficial to present the listener with a process which familiarizes them with the range of test material they will hear. Bech suggests this process be achieved through a method he refers to as *sample familiarization*. In this method "both the sample and user-interface familiarization can be combined into a training experiment, where the subjects perform a subset of the experiment with a set of samples that span the range of qualities to be evaluated. In this manner, they hear the range of samples and can also become familiar with scaling their perception with the UI" [91]. Bech goes on to say that this process, in his experience, leads to results that are more reliable.

2.5 Mixing for Picture

Parts of this thesis focus on auditory motion in the presence of real-time visuals. This section will provide relevant background on mixing for picture, including common loudspeaker/sound source positioning formats, recommendations, and basic aspects of cross-modality. Further review is suggested through Holman [92].

2.5.1 Stereo and Surround Sound Formats for Picture

An understanding of stereophonic sound (stereo) principles is necessary when mixing for nearly all modern audio productions, including productions in surround sound and binaural. In order to achieve binaural synthesis, for example, the stereo format is used to deliver the two independent binaural audio channels to the ears. Eargle defines stereo as "any system of recording or sound transmission using multiple microphones and loudspeakers. Signals picked up by the microphones are routed to loudspeakers that are located in a geometrical array corresponding to the microphone array. Stereo need not be limited to two transmission channels [12]" although, it should be noted that the final output of a stereo system typically resolves in a Left and Right (single or grouped) channel configuration.

The format was first developed and exhibited in the late 1800s by Clément Ader [93] but modern stereophonic sound is rooted in Blumlein's work in the 1930s [94], including methods for playback and recording, as well as surround sound. It is now widely accepted as the industry standard for most music, film and broadcast productions. Surround sound, like stereo, is a multichannel format which was first developed for the cinema. Unlike stereo, additional channels exist beyond the

Left/Right configuration to provide discrete signal separation or added ambience to a mix [12].

2.5.1.1 Stereophonic Positioning & LCR

The most common application of stereo consists of two loudspeakers arranged at ± 30º to the left and right of the listening position [95] [96]. In large cinemas, multichannel stereo is extended to the sides of the theater, presenting multiplications of the Left and Right channels to each side. In any stereophonic setup with just two speakers, a center sound image is always presented as a phantom image between the Left and Right loudspeakers. For greater spatial resolution in 2-channel stereo systems, specifically in regards to the center image, a Center loudspeaker position may be added to create the LCR (*left, center, right*) format [97]. The Center loudspeaker in LCR is positioned at 0º in front of the listening position, with the Left and Right channels still at ± 30º to the listening position, as shown in Figure 14. It is important to note that a two-channel stereo system must rely on phantom imaging to present a center positioned sound source. The addition of a discrete center channel thus removes potential problems with localization blur. For key elements of a mix like dialog, the engineer can trust that the discrete audio produced through the center channel will localize to roughly the same central screen position for all listeners in cinemas and virtual surround productions (see Figure 14) [12].

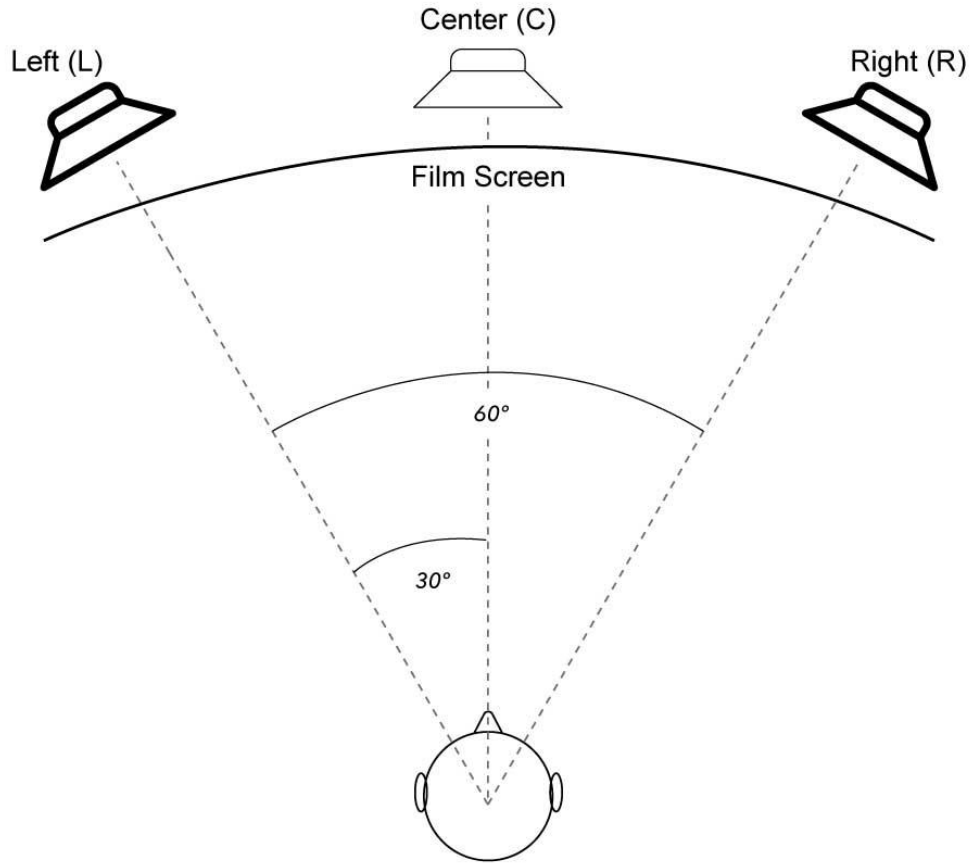

Figure 14: Stereo loudspeaker layout with added Center channel to show LCR

2.5.1.2 Surround Positioning

Surround playback originated in the cinema in the 1950s in the form of a simple four-channel format (LCR plus one surround channel) where the surround channel was emitted "towards the rear of the cinema. At first [the surround channel] was known as the effects channel, and was reserved for the occasional dramatic effect [98]." The most common modern application of

surround sound is the 5.1 channel format, used not only in cinema, but also in high fidelity surround music recordings. In the 5.1 format, there are five discrete audio channels and outputs which include LCR as mentioned previously, plus two additional rear surround channels called *Left Surround* (Ls) and *Right Surround* (Rs). Today, these surround channels are still typically used for ambience and effects. Finally, a *low frequency effects* (LFE) channel is provided for added low-end support via subwoofer [96]. All five full-program channels in the 5.1 format are located on the horizontal plane.

In all surround formats, the number before the point (.) represents the number of discrete program channels and the number after the point (.) represents the number of discrete LFE channels. Today, the ultimate goal of surround sound in the cinema is to achieve spatial audio playback that completely envelopes the listener from all directions (including above and below) and provides the ability to localize sound from anywhere on the screen [99]. To achieve this, a wide range of surround formats exists including 7.1, 10.2 [96], 22.2 [99] and *Object Based* formats like Dolby Atmos [100]. Despite the differences in channel count, the goal in all of these 3D audio surround formats is to present the listener with added immersion greater than 5.1 can provide. For example, support can be added to a 5.1 layout to provide additional discrete side channels through *Left and Right side surrounds* (Lss, Rss), thus creating the standard 7.1 system. These surround channels can be positioned above the horizontal plane to provide a sense of height. Additionally, that same 7.1 format can be altered to relocate the side surrounds to one position *directly overhead* (Tc) and one directly *behind center* (BC) of the listener. In the 10.2 format, an additional height level

consisting of 3 channels is added to the horizontal 7.1 layout, plus the addition of stereo subwoofers positioned to the left and right of the screen.

2.5.1.2.1 Hemispherical Loudspeaker Arrays

It is common practice to use a hemispherical loudspeaker array when measuring HRTFs or presenting 3D audio playback for spatial audio and localization experiments. Therefore, a general understanding of hemispherical loudspeaker surround formats is required. The final experiment of this thesis was performed using a 3D audio playback array derived from the previously mentioned 22.2 surround format developed by *Japan Broadcasting Corporation* (NHK) [99]. The 22.2 format is an extreme version of surround, adding the ability to present sound emitted from "all directions surrounding the viewing position [99]" through a hemispherical array. At such great resolution, it gives the users a "sense of sound approach[ing] from above or below."

Other relevant formats for creating 3D audio projection, including *wave-field synthesis* [101] and *ambisonics* [60] are considered to be out of the scope of this thesis and generally understood by the reader. Further reading is recommended on these topics and their applications through [98] [97] [102] [103] [104].

2.5.2 Cross Modality and Sound Field Positioning

Most of the time, what you see on a film screen is also what you should hear. As Chion & Gorban [105] state, it is often that the viewer of a film "agrees to forget that sound is coming from loudspeakers and picture from the screen." Bregman's work on constructive narrative further

confirms that the sounds we hear in our daily lives are tied to the physical objects from which they originate [11]. Therefore, when mixing sound for picture, it is important to understand that when sound occurs naturally within a scene, a spatial correlation should exist between sound object and visual object. However, this is sometimes not the case, since mixing for picture is as much of an art as it is a science. In some cases a post-production mixer will purposely set the mood of a scene by omitting *diegetic sounds* (sounds that exist naturally in the film scene) for a feeling of suspense. Other times, *non-diegetic sounds* (sounds that don't occur naturally within the scene) are mixed into the scene for added emphasis and effect. In any case, the integration and synchronicity of audio and visual information is an important process to understand in order to control the perceived quality and accuracy of such cross-modal audio-visual events in a given production.

2.5.2.1 Audio-Visual Cross Modality Effects

The perception of one's environment around them is reliant on the ability to process multi-sensory information. As Kohlrausch explains, "auditory and visual stimuli caused by the same event have a specific temporal, spatial and contextual relation when reaching the observer. For instance, when [one observes another] speaking, [there is usually] little difficulty of linking the auditory speech signal to this specific speaker, even if there are numerous other voices present simultaneously [88]." It is also possible to fool the brain into believing that the sound heard belongs to a visual that is not the actual sound emitter. A good example of this is known as the *Ventriloquist Effect* which has been studied at length and focuses on the audio-visual synchronicity of speech. The effect takes its name from the action of the ventriloquist, an event in which the viewer is fooled

into believing (at some level) that the sound of the puppet's voice is actually coming from the puppet's mouth rather than the puppeteer's—essentially moving the localization of the sound source due to visual influence [106] [107]. This skewed perception is due to a dependence on changes in sensory cues within the cross-modal relationship of sight and sound. At certain levels of synchronous processing, the interaction of each sense is processed independently through the evaluation of altered states between the two sensory percepts, which leads to the spatial pairing of sight and sound [106].

2.5.2.2 Temporal Asynchrony in Audio Visual Presentations

When visual objects are in motion, the influence of sound can contribute to cross-modal effects in the perceived motion path of the visual object(s). In separate studies by Sekular and Watanabe [108] [109], a visual presentation of two converging discs was presented to a viewer. The discs traveled along two separate intersecting paths such that the discs would intersect at a point and then continue through each other along their path and onto the opposite side from which they began. Typically, with visual presentation alone, the viewer will observe the discs intersecting, then moving past each other. However, in the presence of audio cues at or near the point of intersection (a range of audio clicks were used in these experiments occurring from ±1000 ms to the intersection event), the discs appear to reverse direction and follow the opposite path. The visual result is that the discs essentially "bounce" off of one another. It was found that sounds proceeding and synchronized with the point of visual object intersection resulted in the bouncing effect more than sounds following the visual intersection. These results signify that temporal

disparity exists during synchronized cross-modal events such that it may be difficult for the brain to retain full attention and process both events equally at the same time.

2.5.2.3 Sound Field Positioning

In general, dialog or other global narrative should be panned Center, music mixed separately is spread between stereo Left and Right, ambiance is spread to the Rear Surrounds, *sound effects* (FX) and *Foley* (live performance FX) are panned to the position of the sound source on screen, and dialog entering the scene from off-screen is panned from the Surrounds towards Center. Considering the studies and points referenced in previous sections, Holman suggests the following standards when mixing sound to film [92]:

- Dialog in ongoing conversations is usually either centered or kept close to center because sound edits that match picture edits cause the sound to noticeably "jump" around the screen
- Off-screen lines are usually panned hard left or right, as makes sense; panning them to the surrounds in the auditorium breaks the "box" of the frame line too much
- Lines that are isolated from others in time may be panned
- Foley is routinely recorded in mono and panned into place to match
- Ambience is most often from original stereo recordings that are placed from two or more source channels into two or more output channels. The principal aesthetic concern of ambience panning is whether to include the surrounds, depending on whether the audience is supposed to be sharing the space portrayed on the screen

3 Lateral Listener Movement on the Horizontal Plane: Sensing Motion through Binaural Simulation

Abstract

An experiment was conducted to better understand first-person motion as perceived by a listener when moving between two virtual sound sources in a VAE. The question was asked: if auditory motion was simulated between two sound source positions using binaural crossfading, how would a listener rate the overall motion they perceived, herein termed "sensation of motion", if it was compared to the same stimuli without motion and to a third, reference binaural recording of the actual motion path? A motion apparatus was designed to move a HATS between two matched loudspeaker locations while recording various stimulus signals (music, pink noise, and speech) within a semi-anechoic chamber. Synchronized simulations were then created and referenced to video. In two separate, double blind MUSHRA-style listening tests (with and without visual

reference), 61 trained binaural listeners evaluated the sensation of motion among real and simulated conditions. Results showed that the listeners rated the simulation as presenting the greatest sensation of motion among all test conditions.

3.1 Introduction

At the time of this publication, virtual reality has recently become a commercially assessable product, and not just a developmental technology limited to the gaming industry and three-dimensional (3D) audio and video research. As such, virtual reality requires the delivery of high quality, realistic three-dimensional audio processing of VAEs to facilitate listener immersion in a virtual 3D world. More than 10 years have passed since Blauert foresaw the use of such applications in VAEs being interactive individualized movie sound, tele-conferencing, virtual sound studios and more [110]. However, despite 10 years of research since then, the deliverable product of 3D audio is not quite what most audio engineers would call “high quality.” One factor in particular that seems to affect this perception is 3D auditory motion. As virtual reality mainly focuses on binaural delivery of 3D sound, this paper will focus on the issues related to headphone perception of 3D auditory motion.

Several methods have been proposed and/or evaluated in the literature for the simulation of binaural motion. For example, methods exist for: a motion-tracked binaural (MTB) sound apparatus with microphones placed on a sphere to capture relative, dynamic proximity of sound sources at each ear [111]; the interpolation of HRTFs based on IPTFs [68]; and direct switching of HRTFs [67]. Matsumoto et al reported “methods to interpolate binaural impulse responses” [75]

wherein a “simple method” of binaural motion simulation, referred to as the “conventional technique” was described. This conventional technique for crossfading binaural impulses through convolution used non-dynamic linear interpolation and therefore, is viewed by the author as the most basic method of auditory motion simulation, since it does not represent the “arrival time difference of the two responses.” [75]. Informal analysis of this technique implies that it might perform best for stationary listener positions to present moving sound images, and may not necessarily represent an auditory motion path through space from the listener’s perspective, but rather sound objects moving around the listener’s head.

The previously published research by Matsumoto et al [75] begins to study the perception of self-motion through binaural motion simulation. The authors investigated a custom algorithm for binaural motion using HRTF convolution and time arrival interpolation. The motion captured in the experiment was achieved through rotation of a dummy head at a fixed position within a free field. The dummy head therefore did not traverse in space, but exhibited rotational listener motion. This algorithm was evaluated against an actual moving dummy head recording and the “conventional” binaural crossfade. The results of their experiment showed that the algorithm was evaluated at ratings near that of the actual recording.

As we begin to discuss auditory motion, one must consider the primary auditory cues related to this perception. As sound moves in space from a specific location and reaches the ear, there will be an arrival time and audible level associated with that path. Since the head separates the left ear and right ear, any difference in arrival time and audible level at the individual ear positions is

denoted as the Inter-aural Time Difference (ITD) and Inter-aural Level Difference (ILD), respectfully [110]. When dealing with two or more simultaneous sound source positions, ITD and ILD values will be present at each ear for each individual sound position [110]. This situation becomes very difficult to manage in simulations because constant motion requires constant variation of ITD and ILD values for each sound source. Thus, when a listener moves in space between sound objects, the situation becomes even more complicated [110] [112].

3.2 Background

3.2.1 Auditory Cues on Motion Perception

The relationship of changes in binaural cues (ITD and ILD) to auditory motion perception is presented in a publication by Kapralos et al [112]. In the same publication, Kapralos et al created an experiment to understand physical motion and auditory motion as separate and combined factors on the perception of self-motion. It was found that auditory cues were highly significant to the accuracy of perceived self-motion. Moreover, when combined with physical motion, auditory cues improved the accuracy of self-motion perception. As such, the primary author modified this previous experiment to include virtual motion simulation through the capture of physical motion, to answer the needs of the following hypotheses.

As 1st person virtual reality applications move the listener position within a VAE, i.e., in the case of a binaural walk-through; motion perception should become a factor in the success of the presented auditory stimulus based on accurate localization and the presentation of auditory cues.

It was of interest whether this motion could be simulated through a conventional binaural crossfading technique, and if this technique could provide significant auditory cues for the perception of motion. Furthermore, would this perception equal that of an actual binaural motion recording?

The following hypotheses were tested:

1. It is assumed that the technique of binaural crossfading should be an efficient method for simulating auditory motion on the horizontal plane between 2 separate sound source locations.
2. It is assumed that the overall motion phenomena perceived by listeners ("sensation of motion") in the auditory motion simulations will be rated similarly to the evaluation of a reference motion recording.
3. It is assumed that the presence of visuals and acoustic reflections will influence the sensation of motion ratings.

To reduce variables and perform an experiment with controlled parameters, an investigation into this perception was performed for a listener moving laterally between two static sound sources in a VAE. This paper evaluates the *"sensation of motion"* as a metric of perceived auditory motion. This term may also be referred to as *"motion sensation"* within this paper.

3.3 Preparation

An experiment was planned to test the hypotheses of this chapter. The following goals were conceived for the success of the research:

- Reproduce three separate high definition audio stimulus signals of stereo music (jazz quintet), mono pink noise and mono speech through two matched loudspeakers
- Perform the experiment in a controlled acoustic environment for the capture of multiple recorded stimuli, thus limiting any influence of room reflections
- Capture moving recordings of high definition video and binaural audio, which travel horizontally from one loudspeaker to the other on a lateral plane positioned in front of the loudspeaker positions
- Create a visual tracking system to synchronize the experiment's audio, video and simulations
- Gather BRIRs at multiple listener positions for simulation
- Create binaural crossfade motion simulations
- Develop a localization training test for listening test participants
- Develop two separate listening tests (with and without video reference) to allow for the perceptual evaluation of motion sensation for recorded and simulated stimuli
- Record all test data to demonstrate significance for multiple factors by way of parametric and/or non-parametric statistical analysis

3.3.1 Reducing Experimental Bias

In this experiment, the goal was to capture and reproduce audible motion with the fewest variables so to reduce experimental bias due to unintended variables. Physical motion, rather than a computer simulation, was chosen so that a real-world example of motion perception would exist and serve as the experimental reference for the independent variable: the binaural auditory motion simulation. Since the operation of physical motion itself introduces many variables such as noise and physical vibration, extreme precision and care were performed at all stages of the experiment. Any inconsistencies in the method of replicating the reference signal might introduce signal

colorations that would potentially bias perceptual evaluations. This happens when a listener's responses are unconsciously weighted through a process of sequential contraction whereby the order of presented stimuli may bias listener responses toward previous or subsequent stimuli should significant changes in signal attributes occur [113]. However, it should be noted that all simulations were created from the actual binaural audio recordings, which would include any such variables as well and thus, negate their effects on the perceptually compared results.

3.4 Measurement

Research was conducted at the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) at McGill University in Montreal, Canada using high quality, measurement grade labs and equipment. As the experiment required little to no acoustic reflections, the semi-anechoic chamber was chosen to facilitate the experiment and test measurement (see Figure 15) with total room volume of 124 m^3 and a reverb time (T30) of 90 ms. Since a fully anechoic lab was not accessible during this study, the semi-anechoic chamber was chosen to provide an environment which greatly limited the influence of acoustic room reflections on recorded and simulated binaural stimuli. The experiment also required the design of a low-noise motion apparatus, as well as the capture of high definition video, binaural audio and BRIRs for later use in simulating the real binaural auditory motion. In order to accurately simulate the motion of the binaural recordings, positional references were required for synchronization purposes.

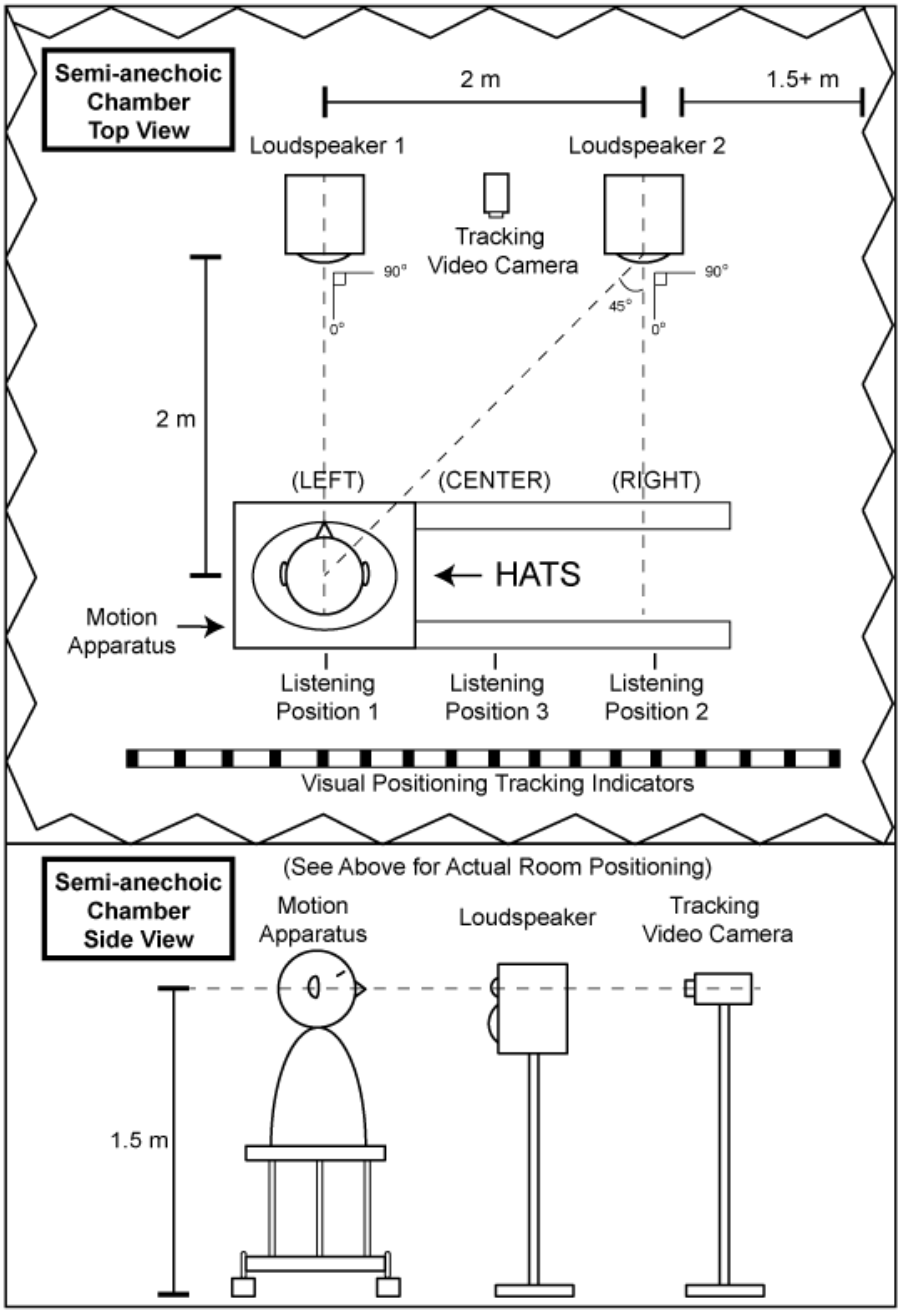

Figure 15: Layout of the semi-anechoic chamber used for experimental measurement

Inside the semi-anechoic chamber, a matched pair of full range, high quality loudspeakers was

placed at an equal height of 1.5 meters and a distance of 2 meters apart (see Figure 15 and Figure 16). Both loudspeakers were positioned to face listening positions 1 & 2 for sound reproduction at an elevation and azimuth of 0 degrees. A motion apparatus (Section 3.4.3) was placed horizontally in front of, and facing the speakers at a distance of 2 meters with the ends of the apparatus aligned to present a listening position directly in front of each loudspeaker location. Mounted on the motion apparatus was the HATS, aligned to an elevation and azimuth of 0 degrees to the horizontal plane of the loudspeakers. The ear canal height of the HATS was aligned to the loudspeaker height of 1.5 meters from the ground. The entire setup maintained a distance of 1.5 meters or greater from the semi-anechoic chamber walls.

3.4.1 Measurement Calibration

Both loudspeakers were calibrated to have a flat frequency response of 35Hz to 40kHz using measurement standard calibration microphones and equipment for loudspeaker equalization based on the specifications defined in a publication by the Army Research Laboratory in Maryland [114]. A digital measurement system capable of recording and analyzing high definition audio impulses, reproduced and recorded all calibration audio at 24 bits/96 kHz. Loudspeakers were individually calibrated for an output of 90 dB SPL (*Sound Pressure Level* - C weighted) at the HATS position so as to provide a significant signal to noise ratio.

It was also necessary to ensure accurate sound arrival at each ear in addition to dimensional positioning measurements mentioned previously. Therefore, at each extent of the apparatus (see position 1 & 2 in Figure 15), the corresponding loudspeaker was used to provide stimulus signals

at 0 degrees for measurement of ITDs and ILDs between both ears on the HATS. The measurements were aligned to within an ITD less than 0.01 ms and an ILD less than 0.046 *decibel units* (dBu).

3.4.2 Reproduction stimuli

To begin the measurement process, three high-definition audio signals were chosen to become the "reproduction stimuli" for the binaural recordings. The three stimulus signals were loudness matched to the ITU 1770-3 recommendations [115], therefore reducing bias in the recordings due to loudness differences. The reproduction stimuli included a stereo live music recording (jazz quintet), mono pink noise, and mono anechoic male speech [116]. Broadband signals were included as stimuli since they have been determined to display more accuracy in sound source distance perception than single tones [117]. All signals ranged from 10-12 seconds in duration and were reproduced through both loudspeakers equally, except for the stereo music, which was presented in stereo across the two loudspeakers. This decision was made so to present a wider range of stimuli for evaluation. Trained professional audio engineers listened to the reproduction stimuli through the loudspeakers in the semi-anechoic chamber, and after several rounds of listening, it was determined that live music provided a realistic sound stage while pink noise provided full broadband evaluation and speech provided a familiar, individual sound object at a location in space.

3.4.3 Motion Apparatus

Great care was taken during the design and construction of the motion apparatus, as it needed to

provide accurate and consistent movement through the experimental setup.

3.4.3.1 Apparatus Design

The apparatus included a rolling base with a mounted platform—measured for accurate placement and control of the HATS during operation. Low friction wheels and suspension were added to absorb vibration, while steady motion was provided by a rope and pulley system on a position locked track. Audio cabling from the HATS could move freely within the apparatus so as to not induce additional noise or signal disturbances. Position in space and time was tracked by the addition of visual markers hung behind the apparatus as captured by a stationary video camera (see Figure 15). Points of reference through the video camera were then marked in a post-measurement review of the video to ensure proper positioning for later simulations. Finally, the apparatus was fitted with a video camera mount to also track visual motion for the listener point of view, which matched the audible recordings made during each round of motion recording. The level of operable noise was extremely low for the entire system and undetectable in the binaural recordings, thus treated as an insignificant value in the overall signal to noise ratio of the audio recordings.

3.5 Procedure

The experiment required a two-stage procedure to be performed. The first procedure captured high-definition BRIRs for simulation whereas the second procedure captured the real-world binaural auditory motion with synchronized video references.

3.5.1 Binaural Auditory Motion Simulation

Once the apparatus and loudspeakers were calibrated and aligned within the semi-anechoic chamber, BRIRs were collected through the HATS at each of the three listening positions located on the horizontal plane of the loudspeakers (see Figure 15). Therefore, an HRTF representing the spatial positioning information of a single BRIR could be used to simulate motion by crossfading its convolved audio with that of another HRTF representing a separate, virtual location [111] [68] [75]. It then became possible to virtually produce sound localization to the listener for two simultaneous sounds, each appearing from three possible locations in front of the listener at 0 degrees elevation and azimuth angles of (see Figure 15):

- 0 degrees center and 45 degrees right (pos. 1)
- 0 degrees center and 45 degrees left (pos. 2)
- 26.6 degrees left and 26.6 degrees right (pos. 3)

Note that 26.6 degrees is the angular azimuth at the rear center of a 2 by 2 meter square. Since localization blur causes less differentiation in auditory space for a sound source as angular displacement increases from the forward position [14], the angles of the sound source locations for this experiment were chosen to reduce the effect of localization blur while providing sufficient localization differences in the audible stimuli.

3.5.1.1 Detail of Impulse Response Design & Binaural Signal Convolution

To present simulated binaural cues for localization and auditory motion experiments, the common method is to convolve test signals with a pair of HRIRs to reduce the influence of room reflections.

This process prevents unwanted signal coloration during binaural simulation and allows the researcher to focus solely on the perception of binaural cues within a free-field. However, in this experiment, it was essential to represent all of the acoustic energy that would be present through the reference binaural real-world recording of auditory motion within the semi-anechoic environment. The test stimuli would need to be convolved with the full binaural acoustic response of the room, and not just the early HRIR portion.

Using impulse response measurement software, both equally calibrated loudspeakers reproduced a single logarithmic sine sweep at 90 dB SPL (C weighted) from 20Hz to 20kHz at 24 bits/96kHz, over 10 seconds. Using one loudspeaker at a time, the measurement software recorded the sweep through both ears of the HATS at a given position. This process generated a 2-channel binaural recording (Left Ear/Right Ear) of the sweep emitted by the Left loudspeaker from the listener's perspective at Position 1. The process was repeated at Position 2 and Position 3 so that 3 separate positional pairs of the Left loudspeaker binaural sweep recordings were made. The process was then repeated for the Right loudspeaker creating 3 more pairs of binaural sweep recordings.

It is important to note that the onset delay from each loudspeaker to its corresponding 0º azimuth position (Left Speaker to Position 1 or Right Speaker to Position 2) resulted in a pre-delay of 6.13 ms ± 0.01 ms (2.10 m ± 0.01 m). At Position 3, this pre-delay was measured at 6.72 ms ± 0.01 ms (2.30 m ± 0.01 m) to the ear nearest to the loudspeaker (Left Speaker to Left Ear, Right Speaker to Right Ear).

The final binaural sweep recordings were deconvolved (per channel) with the single-channel stimulus sweep using a rectangular window set to 97 ms to create final *binaural impulse responses* (BIRs) [118] [119]. It is important that these BIRs are properly referred to as BRIRs since the specific window was deliberately chosen to include the onset signal pre-delay as well as the full acoustic response defined by the measured *reverb time* (time for acoustic energy to decay by 60 dB once the sweep signal is turned off) based on the T30 slope.

Six separate BRIRs were created in total through the final deconvolution process. The BRIRs were labeled as (P1LL, P1LR), (P1RL, P1RR), (P2LL, P2LR), (P2RL, P2RR), (P3LL, P3LR), and (P3RL, P3RR) where P1LL indicates the Position 1 measurement of the Left Loudspeaker through the Left Ear of the HATS. BRIRs were then peak normalized (as pairs) to 0 dBFS (*decibels full-scale*) retaining the proportional intensity relationship between the left and right ears for each BRIR. The greatest intensity in each BRIR was present in its HRIR portion, therefore normalization accurately retained the onset binaural cues.

Full detail of the BRIRs is presented in Figure 36 through Figure 41 indicating each HIRIR pair, along with the additional acoustic influence of the semi-anechoic environment on the binaural measurements. Upon close review, floor reflections are present at roughly 3.5 ms following the onset signal. A significant reflection at around 24 ms can also be seen in the graphs of the full BRIRs (see Figure 36, Figure 38 and Figure 39). This reflection is presumed to come from the entrance door (located on the back wall between Positions 1 and 3) since the full reflective path would have been roughly 8.2 m (accurate for the position measurements which include this

reflection). Such reflections from the floor and the door would significantly color the convolved stimuli and if left out from the simulation process, would have significantly reduced the ability for the simulation to accurately replicate the binaural reference recordings.

Once BRIRs were generated, the stimulus signals (Music, Pink Noise, Speech) were then separately convolved through each of the six BRIRs using signal convolution software. This process created a rendered binaural simulation signal representing each loudspeaker at each listening position for all 3 stimuli. To complete the simulation stimuli, which would emulate simultaneous playback from both loudspeakers as heard from each position, the left loudspeaker binaural stimulus signal and right loudspeaker stimulus signal were summed for each position. This created 3 final, position specific binaural simulation stimuli. These binaural simulation stimuli would be crossfaded to deliver the final binaural auditory motion simulations.

3.5.1.2 Crossfading Technique for Binaural Auditory Motion Simulations

Through a basic technique derived from previous methods within the literature [68] [75] [120], the BRIR convolutions from each position were binaurally crossfaded so as to be synchronized to the video track of the actual binaural recording made by the HATS. This process created two-channel, binaural audio motion simulation files for all reproduction stimuli. These files matched the motion of the individual binaural audio motion recordings mentioned in Section 3.5.2.

A total of 9 positional markers (see blue strips in Figure 15 and Figure 16) located behind the motion apparatus were positioned equidistant from each other and aligned from the center of the

motion path to the path extents. For a given motion pass, the video could be referenced to the consistent BPM of the recording, but also to the exact time in milliseconds it took the apparatus to travel from one positional marker to the next. Each reference binaural motion recording was matched to the video to extract the timing of motion between all positional markers.

The process of binaural crossfading was performed by linearly crossfading the output gain of one stationary binaural simulation stimulus to another along the motion path. For example, to simulate full left-to-right translational path motion, a linear crossfade was applied to the output gain of Position 1 (beginning) and Position 3 (center). Then as the full output level was reached for Position 3, a second linear crossfade was applied to output gain of Position 3 and Position 2 (end). The ramp time of each crossfade was broken down into time intervals which replicated the same intervals from each positional marker in the reference video for a given stimulus recording. Five positional markers divided the full time period of the translation motion from one position to the next. By associating each position with a percentage (indicated in Figure 42) (i.e., marker 1 = 100% Position 1, marker 3 = 50% of Position 1 and 50% of Position 3, marker 5 = 100% Position 3), the linear crossfade was ramped to reach the signal ratio by the exact time it took the HATS to reach the associated marker in the video. A complete listing of ramp timing is presented in Table 7.

After accurately replicating the timing of the crossfades to the timing of the reference binaural audio motion recordings, the crossfaded binaural audio simulation stimuli were rendered out as a final binaural audio motion simulations for each stimulus and path direction (Left to Right and

Right to Left). As a final step for the comparison of simulations to reference recordings, all binaural audio motion simulations were loudness matched to the respective reference recordings following the standards of ITU-R BS.1770-4 [115]. Several trained audio engineers confirmed accurate sound positioning through an informal listening review over test headphones.

3.5.1.3 Angular Localization Capture

Through the BRIRs captured in Section 3.5.1, five different angular locations along the apparatus track could be represented through HRTF convolution. These positions were used to convolve binaural audio samples of each of the reproduction stimuli, creating as a result, the "localization stimuli." These audio samples were used as auditory stimuli in the localization pre-test (see Section 3.6). They demonstrated five audible positions of an angular localization:

- Center (0 degrees)
- Left Center (26.6 degrees left of center)
- Left (45 degrees left of center)
- Right Center (26.6 degrees right of center)
- Right (45 degrees right of center)

3.5.2 Reference Capture: Binaural Auditory Motion

To capture a reference for the actual binaural auditory motion, the HATS would need to move along the apparatus while recording one of the three reproduction stimuli mentioned in Section 3.4.2. During playback of the reproduction stimuli, binaural audio signals were recorded at 24 bits/96kHz to a *digital audio workstation* (DAW) with high quality audio preamplifiers and

converters. A click track was synced to the stimulus signals within the DAW, which operated playback at a constant tempo. A pre-roll initiated the synchronization playback and motion of the apparatus.

To begin the capturing process, playback was primed in the DAW. As the reproduction stimulus playback began, the HATS recording apparatus was moved at a consistent rate between position 1 and position 2 over the entire duration of one stimulus signal reproduction. During this time, headphone monitoring of the tempo and audio ensured steady motion of the apparatus. Visual motion of the HATS as it travelled along the apparatus was captured by the central position video camera (see Figure 15 & Figure 16).

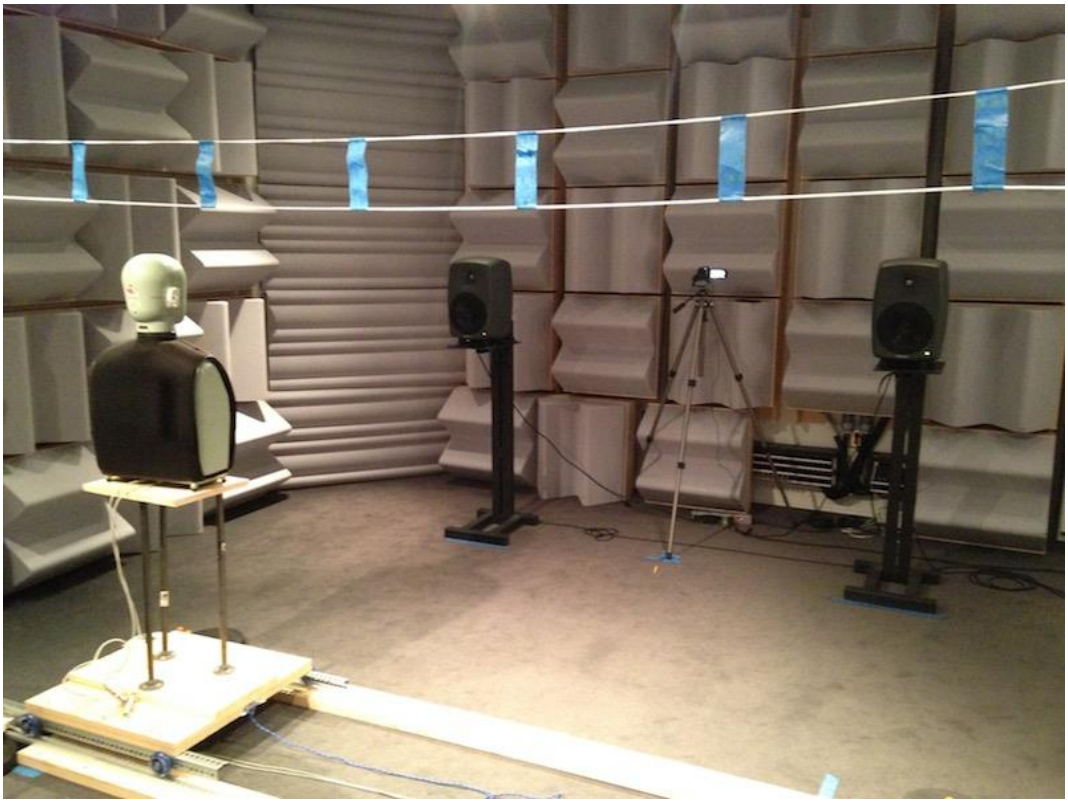

Figure 16: Image of the semi-anechoic chamber showing the HATS, full motion apparatus, loudspeaker positioning and video tracking system

This process was repeated several times to provide the best average accuracy recording. After the recording of one stimulus was completed from position 1 to position 2, the same process was repeated in reverse from position 2 to position 1 to achieve recordings of both, left to right and right to left direction on the horizontal plane. After completing the reverse process, the next reproduction stimulus recording was made until all passes of stimuli/direction were recorded. All binaural motion recordings were labeled with a naming convention to effectively organize video recordings to binaural audio recordings. For example, a video or audio file with the name "*L2R-*

120-Pink-L-T1" represented the paired video and binaural audio recording for "L2R" (direction of motion from Left to Right (Position 1 to Position 2)), "120" (tempo in beats per minute (BPM) to which the velocity of the motion apparatus was synchronized), "Pink" (the pink noise stimulus signal), "L" (the Left loudspeaker from which the stimulus was reproduced), and "T1" (trial recording #1). With this organized method, a video could be easily recalled within the DAW along with the associated (paired) binaural audio motion recording during the final crossfading process for binaural auditory motion simulation. Each motion path was roughly 8 seconds in duration. The greatest variance in path duration between a stimulus motion path (L2R and R2L) was 290 ms for the Speech stimulus. All others fell below 200 ms. Finally, a video camera mounted to the apparatus, repeated the procedure for each situation. This process captured the listener's visual, 1st person point of view for reference in the listening test.

3.6 Listening Test

Two types of listening tests were developed to critically evaluate the audio gathered in the experiment. MaxMSP was used to create a localization training pre-test (see Figure 17) along with two versions of an audible movement listening test (see Figure 18) to accurately evaluate the sensation of motion perceived by the listener when presented with simulated and real binaural audio recordings. The participant pool included 61 moderately experienced [14] [121] participants. These participants were given randomized assessor numbers (automated by the test software) and then asked to fill out a survey for background information pertinent to the research. These participants ranged in age from 18 to 37 and had one half year or more of formal audio training.

The population was made up of 21 females and 40 males.

3.6.1 Listening Room

All listening tests were performed in a soundproof listening suite at CIRMMT. Each MaxMSP-hosted test was performed on identical equipment with high-quality, professional converters and headphones designed for critical listening of audio. Each pair of headphones was calibrated to a level of 75 dB using pink noise while both audio systems were aligned equally. The listening room was divided into two partitions. On one side of the partition, a table was setup to host the audio only version of the final test while on the other side of the partition; a table and large high definition video screen were setup for the audio/video version of the final test. For the audio/video version, the loudspeaker height, along with the distance from the loudspeaker to the listener were mimicked to match the visual point of reference from the measurement stage. This was also in accordance to requirements for the "reference seeing distance" [122].

3.6.2 Localization Training Pre-Test

Since all test signals, including simulations, stemmed from measurements made with a head and torso simulator instead of the test participants' individualized ears, it was essential to determine if a large difference existed among participants on localization accuracy for the HRTFs used in the creation of the sound source simulations [14] [123]. Otherwise, error in localization could cause bias in the results of the audible movement listening test. Therefore, a threshold of 90% total average localization accuracy was set to determine successful localization. This localization assessment also served as a training session for familiarization with the scope and range of the

stimuli following Bech's suggestions on sample familiarization [91], as well as the software and controls used in the audible movement listening test.

The localization training pre-test (see Figure 17) consisted of a double-blind, multi-stimulus test influenced by the *Visual Analog Scale* (VAS), *Visual Sort and Rate* (VSR), and forced paired-comparison methods [124] [91]. It asked the participant to localize each audible example (the binaural audio files processed in Section 3.5.1.3) to the nearest position, as demonstrated by an on-screen visual diagram. The diagram contained object locations for which the participant used to deduce his or her answer. The answer was then provided by way of a method influenced by the ACR scale [14] [78], denoting the referenced positions (see Figure 17). Through six randomized trials, one of the three stimuli was presented over five randomized audio examples (A-E). This process was repeated at random so that each localization stimulus signal was presented twice in the pre-test. Comments were optional for each trial.

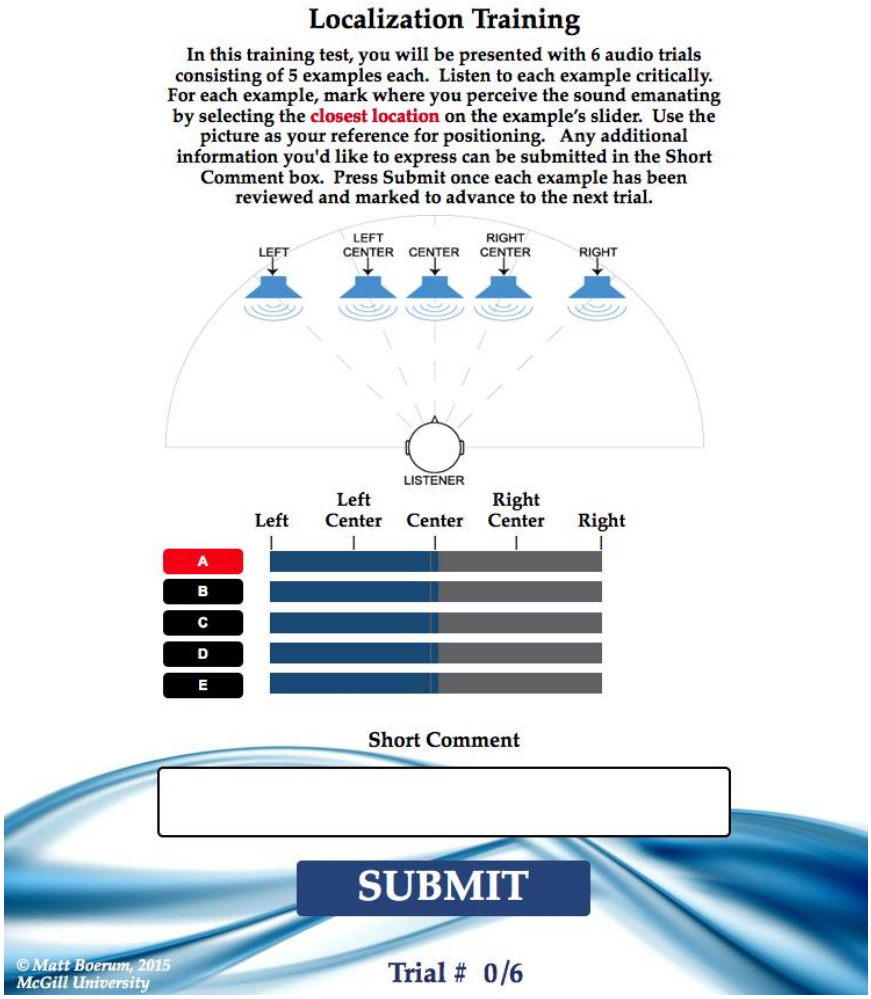

Figure 17: Localization pre-test

3.6.3 Audible Movement Listening Test

Immediately following the localization pre-test, participants were given instructions for the audible movement listening test (see Figure 18). As mentioned earlier, two versions of the audible movement listening test were created—the audio only version and the audio/video version. Participants were selected at random before starting the localization pre-test to determine which

final test version they would take. A total of 31 participated in the audio only version while 30 participated in the audio/video version of the test. These different versions meant that the presence of video could be evaluated as a factor on auditory motion perception.

The audible movement listening test consisted of a MUSHRA-style [77] evaluation of the independent variable—the binaural auditory motion simulations. This meant that the actual binaural auditory motion could serve as a hidden reference while a static BRIR convolution, containing no motion would serve as the anchor. By presenting the participant with a CQS [77], data could then be analyzed on a consistent basis to that of the hidden reference. The scale consisted of 100 points with five quality descriptors ranging from “Bad” to “Excellent.” This proved to be successful and gave interesting results as discussed in the following section.

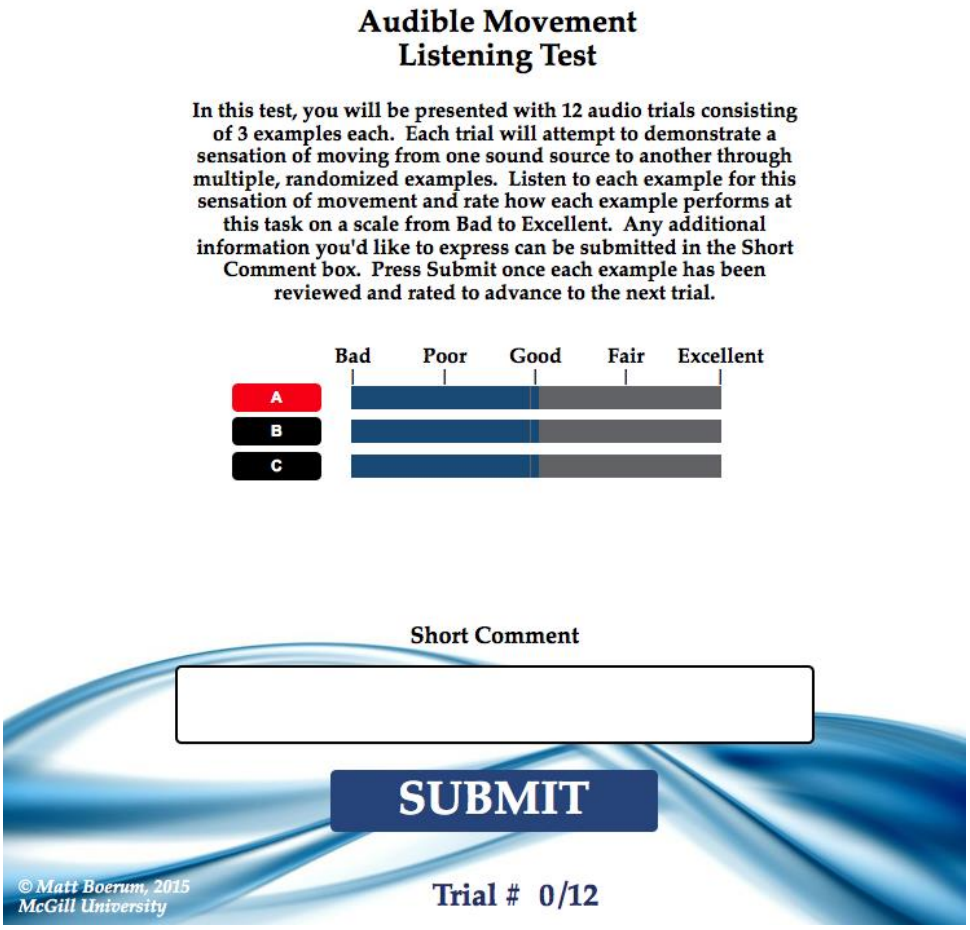

Figure 18: Audible movement listening test

3.6.4 Test Presentation

The test presentation was similar in operation to that of the localization pre-test. However, in both versions, the test presented the participant with 12 randomized trials. One stimulus signal (music, pink noise, or speech) was chosen at random, for each trial and presented as three separate audible examples (A-C). These examples played back randomized conditions (actual binaural motion, simulated motion, or static convolution) of the chosen stimulus signal while demonstrating one of

two directions (left to right or right to left). Therefore, in each trial, a participant evaluated binaural auditory motion of a specific direction and stimulus for the actual HATS recording, the binaural crossfading simulation, and a static convolution. Each stimulus by direction was repeated twice within the test, thus giving 12 total trials. The participant was asked to listen to each example and rate the sensation of motion they perceived from each, based on the CQS [77]. A "bad" sensation correlated to no movement at all, where an "excellent" rating signified a lot of motion. As a guide, participants were instructed to evaluate each audio example for a sensation of motion rather than a specific direction of motion since the experiment was not specifically targeting direction as a factor in motion perception. Therefore, it should be noted that an "excellent" rating only applies to the fact that the participant sensed a great measure of auditory motion and that other factors such as accuracy in motion might have been ignored. All trial examples were infinitely repeatable to account for any perspective-based issues with audio or video unfamiliarity. Comments were optional for each trial.

3.6.4.1 Test Version Differences

Participants were not aware of the different test versions as the room was partitioned. For the audio only version, participants were just presented with audio playback through headphones. In the audio/video version, the exact same test was given but with the addition of the video images recorded during the HATS recordings. As the auditory motion example was reproduced, the synchronized video displayed a listener point of view moving from one loudspeaker to the other in the same direction and velocity as the auditory motion (see Figure 19). There was a point in

time when both loudspeakers were completely out of sight as the example audio approached the central position (between loudspeakers). After moving past the center, the other loudspeaker came into view and ended at the center screen. Participants were instructed to face forward without moving while performing the test, and to focus on the video motion when considering motion sensation. Both versions of the test tracked the duration of each trial by the participants for analysis of test factors by trial duration.

Figure 19: Participant performing the audible movement listening test (audio/video)

3.7 Results

Decisions for statistical analysis and data collection were planned during the preparation stage as mentioned in Section 3.3. The data from the tests could be used to check for significant differences on means through an *analysis of variance* (ANOVA) as referenced by the ITU recommendations for MUSHRA [77]. If the data proved to be non-parametric, the CQS results could be analyzed on ranks by means of the Kruskal- Wallis test (one-way analysis of variance performed on three or more unrelated groups) for non-normal data. Therefore, either outcome would mean that the data could be displayed in several forms including histograms and boxplot diagrams.

The final test data was grouped into five factors of condition, signal, response, trial duration and gender while the localization pre-test data was grouped into four factors of response, accuracy, trial duration and gender. The three factors investigated in both versions of the audible movement listening test were **Condition, Signal and Response**. The three Conditions were **Reference** (the actual binaural motion recording), **Simulation** (the binaural crossfade simulation) and **Anchor** (the static BRIR convolution which contained no motion). The three Signals were **Music** (a stereo recording of a live jazz quintet), **Pink** (mono broadband pink noise) and **Speech** (a mono male speech recording). **Response** refers to the assessor rating value of motion sensation for each presented signal condition on a scale of 0-100.

3.7.1 Data Normality & Significance Testing

To determine normality among all test data groups, histograms were first created and analyzed. Through this analysis, the data proved to be non-parametric.

The Shapiro-Wilk normality test (utilizing the null hypothesis) concluded that the sample data did not come from a normally distributed population for all conditions in both versions of the test described in Section 3.7 (see Table 1). Therefore, the Kruskal-Wallis test was used to perform one-way analysis of variance on the distribution of data. This data would be tested for a p value of $p < .05$. All statistical analysis and graphical plotting were carried out in R Studio and MATLAB.

Table 1: Shapiro-Wilk non-normality values for both audible movement listening tests

Condition	Audio Only		Audio/Video	
	W	**p - value**	**W**	**p - value**
Anchor	.76491	< 0.001	.77146	< 0.001
Simulation	.94633	< 0.001	.94466	< 0.001
Reference	.95995	< 0.001	.94775	< 0.001

3.7.2 Pooled Responses

The assessor responses (ratings) were pooled together for each version of the final test (Audio Only, Audio/Video).

3.7.2.1 Audio Only Data

For the Audio Only responses (dependent variable), there was a statistically significant difference between all comparisons of the signal by condition, $H(2) = 659.1$, $p = 0.047$. Significance was also found between comparisons of the reference to simulation for each individual signal ($p < .002$ separately for music, speech and pink). Median values of the responses for all signals plotted by

condition can be visually inspected by the boxplots in Figure 20. From this statistical analysis, it is important to note that the simulation was rated higher than the reference. However, when comparing each condition by individual signals, the reference scored higher ratings of motion sensation for the speech signal. The anchor ratings proved that the anchor signal processing displayed no motion and served well as an anchor since the ratings were well below the other conditions for all signals.

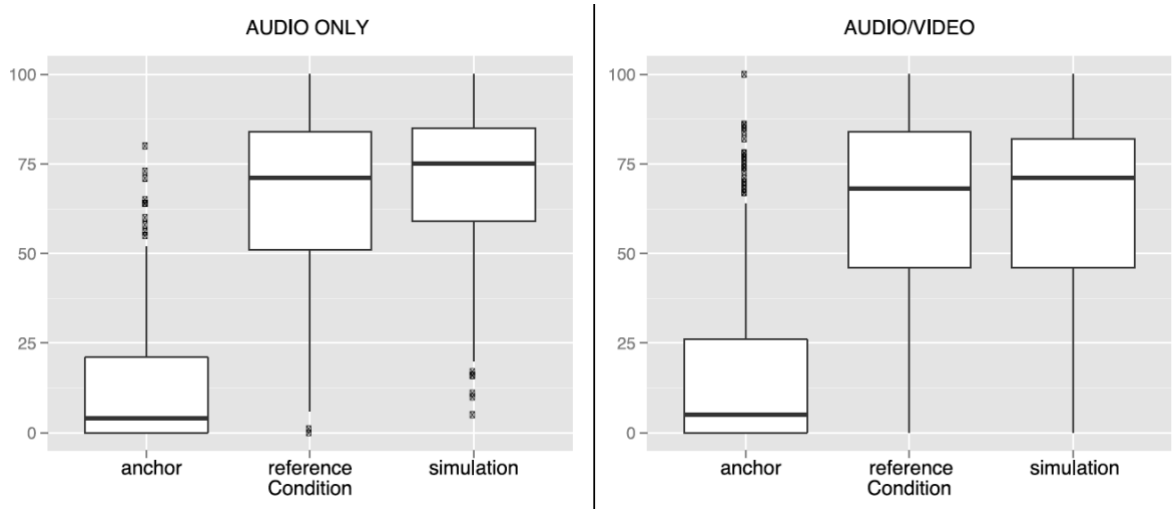

Figure 20: Responses by condition for all signals (from left to right, Audio Only and Audio/Video)

3.7.2.2 Audio/Video Data

Analysis of the Audio/Video version did not find a statistically significant difference ($H(2) = 459.7$, $p = 0.81$) between all comparisons of the signal by condition. This was due to a high p-value between the comparisons of the reference to the simulation for the music signal responses. Evaluation of the boxplot on the right in Figure 20 shows limited difference between simulation

and reference in the Audio/Video version—a visual confirmation as well that these two groups as a whole were not significantly different in this test version for all signal types. Figure 21 demonstrates the difference in ratings for individual signal types by condition for both versions of the test.

Although all comparisons of the signal by condition did not prove to be significantly different in the Audio/Video version, it is still important to note that low p-values existed between the reference and simulation ($p < .004$ for each). Therefore, the speech and pink noise could still be evaluated for the comparison between the reference and simulation. The results showed that the simulation gave more sensation of motion for pink noise, while the reference rated higher again for speech (as was true in the audio only version). The anchor ratings were, as expected, much lower than all other conditions, which was similar to the audio only version.

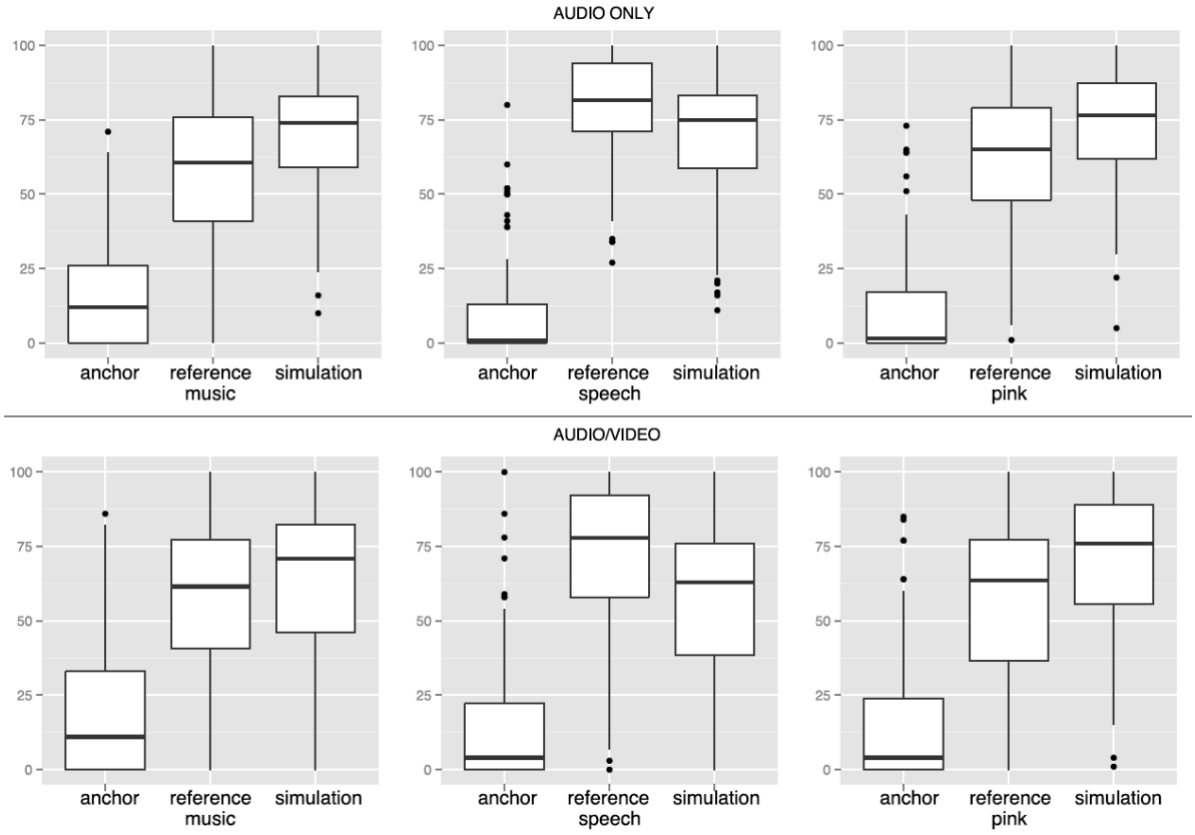

Figure 21: Responses by signal for individual conditions (from top to bottom, Audio Only and Audio/Video)

3.7.2.3 Localization Pre-Test Accuracy

Localization responses were recorded during the localization pre-test to determine the percentage of localization accuracy for each participant. The results were pooled as a whole as there was only one version of the localization pre-test for all participants. The mean accuracy for all participants was 92.7% (see Table 2). Localization error by position is charted in

Table 3, which demonstrates that a significant amount of errors came from the left side. This could be due to experimental error from mismatched BRIR measurements at the left and right angular positions. As will be discussed more in Section 3.8, it is interesting to look at the percentage of error by signal. Speech accounted for 39.1% of all errors, with pink noise at 33.1% and music at 27.8%.

Table 2: Localization accuracy for individual sound source positions, total localization accuracy (Localization Pre-Test)

Sound Position	# of Trials	Accuracy	Errors	Total Accuracy
Left	366	89.1%	40	92.7%
L. Center	366	91.3%	32	
Center	366	96.4%	13	
R. Center	366	91.8%	30	
Right	366	95.1%	18	

Table 3: Localization error by position (Localization Pre-Test)

Sound Position	Total Errors	# Errors by Position				
		L	LC	C	RC	R
Left	40	N/A	39	1	0	0
L. Center	32	5	N/A	26	1	0
Center	13	0	9	N/A	4	0
R. Center	30	0	1	19	N/A	10
Right	18	0	0	0	18	N/A

3.7.3 Gender Specific Results

As the experiment was performed with a HATS modelled after a male head shape, ear shape and torso; it was of interest to investigate any differences existing among gender. A comparison of condition to signal was performed to determine any such differences existing in the ratings of motion sensation (see Figure 22).

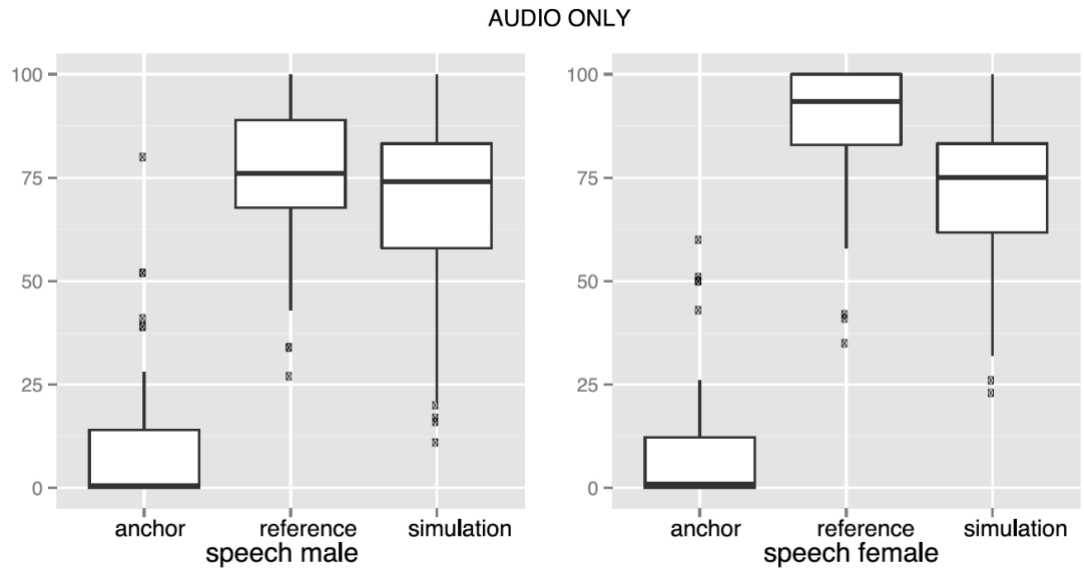

Figure 22: Gender responses for speech by condition (Audio Only)

Figure 22 displays the only significance found among gender related responses. There was a significant increase in motion sensation ratings for "speech" among females to males for the reference condition in the audio only test version.

Trial duration was then investigated and averaged among all participants in both test versions. In analysis of this data by gender, it was found that the mean duration by gender for both versions of the test was greater for female participants than for males. Female participants, on average, spent an additional 9.8 seconds on each trial than their male counterparts.

Finally, localization was investigated by gender. There was no significant difference between the average localization accuracy of females to males in the localization pre-test responses (Audio Only: $H(2) = 2.76$, $p = .736$; Audio/Video: $H(2) = 6.77$, $p = .2381$). The mean accuracy of

localization by gender was 93.7% for females and 92.3 % for males.

3.8 Analysis

3.8.1 Localization Accuracy & Error

The ability for the captured HRTFs to simulate accurate angular localization of virtual sound sources was validated through the high percentage of localization accuracy exhibited among all participants (92.7%). Since these HRTFs were used to convolve the audio for the binaurally crossfaded simulations, it also follows that the ratings for the sensation of motion perceived by the participants through simulation are also valid. Furthermore, the reference recordings of actual binaural auditory motion would also be validated since any real-time recording would serve as the most accurate example of motion simulation using crossfading techniques of binaurally produced HRTF convolution [67].

3.8.2 Sensation of Motion

The simulation results show that the sensation of motion perceived by the listener was at most times greater than that of the actual binaural recording. The one exception was that the reference exhibited higher ratings of motion sensation for the speech signal in both test versions. This is thought to be due to the familiarity of male speech over other test signals [116], such that errors in motion simulation would be more detectable in a familiar signal. Interestingly enough, the comments (see Section 3.8.2.2) regarding the speech signal stated that the simulation seemed to have presented a sense of physical movement which was “contradictory […], off-putting and

uncomfortable." It is unclear whether the assessor experienced this phenomenon through self-motion or apparent motion, however this comment, along with the support of significance testing and box-plot results (see Figure 21) suggests the simulation technique is somewhat deficient in providing motion sensation for male speech. It should also be noted that the anchor displayed little to no sensation of motion as would be expected for a static convolution.

3.8.2.1 Video as a Factor

The presence of video as a factor in the audible movement listening test did not display a significant difference in the results, only a slight proportional reduction in ratings. This is perhaps due to the fact that this experiment asked the participant to focus on non-direction related motion sensation, as mentioned in Section 3.6.3. It should be noted that measures of motion sensation reported through the presentation of music with video were not found to be significantly different and therefore, this individual condition cannot reject the null hypothesis.

3.8.2.2 Listening Test Comments

Comments given by participants demonstrate interesting points for discussion. These comments were given during the listening tests. Here are a few that demonstrate the varying perceptions of motion sensation:

> ***Assessor V016:*** *"The voice definitely gave me a more physical sense of movement. I found the motion of the sound that was contradictory to the motion of the camera really off-putting and uncomfortable."*

Assessor V018: *"Example A (the simulation) felt aligned with the video."*

Assessor A001: *"Example A (simulation) travels from left and right and it creates the sensation of dancing and hearing the band play as you spin around the room."*

3.9 Future Work

This experiment was designed to evaluate binaural motion simulations to actual binaural recordings on the basis of motion sensation alone. As such, further investigation and research into factors such as accuracy of direction, trajectory, positional velocity, and acoustic reflections would be required to create a complete comparison to prove or disprove binaural crossfading as a sufficient equivalent to the human perception of motion through auditory cues. Application of this research through virtual reality headsets would also help illustrate the importance of high-quality spatial audio in the presentation of VAEs.

As an addition to the presented measurements in this paper, an additional dummy head was used in a similar, second round of recordings for future evaluation. Both, the dummy head and HATS were also used in a secondary, live acoustic space to perform the same measurements with the influence of strong acoustic reflections and reverberation. This data was not included in the results due to time constraints and was deemed unnecessary to report for the purpose of this initial study.

As mentioned in the methodology, this experiment binaurally recorded and simulated the music stimulus retaining its stereo output through two loudspeakers while the other (mono) signals were reproduced through the same loudspeakers using dual-mono. It is unclear in the data as to why

music with video was the only stimulus presentation that reported results without a significant difference. It is thought that the differences in the stereo spread of music across the motion path had additional influence on these results when video reference of the loudspeaker positions was presented. An investigation into using dual-mono vs. stereo as the source of binaural motion might provide further insight on these findings.

3.10 Conclusion

The results and analysis from this experiment provided only minimal evidence that would reject the null hypothesis that the sensation of motion experienced during the simulation was in some way similar to the reference. Across all conditions, the sensation of motion provided by binaural crossfading was reported greater on average than recorded binaural auditory motion. However, measures of motion sensation were slightly reduced for male speech signals during simulation as compared to the reference. A possible explanation for this finding seems to come through the accessor's comments which suggest that male speech presented non-uniform motion. Upon further review of all comments, several more were found to support this inconsistent motion transition during speech. None of the comments suggest the motion failed to smoothly transition, but some do suggest inconsistent direction in their paths. For example, the following comment describes this potential phenomenon directly:

> ***Assessor A017:*** *"B (speech simulation) feels like it circles around me."*

Interestingly, a final analysis of all assessors' comments showed that this phenomenon is limited

to the speech signal alone. In stark contrast, comments on music and pink noise motion talk about "smoothness", and even "pleasurable motion."

In addition to male speech, Audio/Video results did not exhibit significant differences between conditions when presenting auditory motion through music. While it appears in the review of Figure 21 that music through simulation presented greater measures of motion sensation than the reference, this analysis could not be confirmed statistically, unlike its Audio Only counterpart.

As a final point, though head shapes and sizes differ significantly among genders, it was found that gender results in this test, as a whole, were not significantly different, barring one exception with male speech. This is thought to be due to an averaging of performance occurring from the delivery of auditory motion through a single HRTF convolution to all participants. The success of the HRTF-based simulation could imply that such motion simulations are acceptable for average populations and that individualized HRTFs might further improve this performance.

4 Lateral Listener Movement on the Horizontal Plane: Part 2 Sensing Motion through binaural Simulation in a Reverberant Environment

Abstract

In a multi-part study, first-person horizontal movement between two virtual sound source locations in a VAE was investigated by evaluating the sensation of motion as perceived by the listener. A binaural crossfading technique simulated this movement while real binaural recordings of motion were made as a reference using a motion apparatus and mounted HATS. Trained listeners evaluated the sensation of motion among real and simulated conditions in two opposite environment-dependent experiments: Part 1 (semi-anechoic), Part 2 (reverberant). Results from Part 2 were proportional to Part 1, despite the presence of strong reflections and significant

reverberation. The simulation provided the greatest sensation of motion again, showing that binaural audio recordings presented lower sensation of motion than that perceived through the simulation.

4.1 Preface

This multi-part study is focusing on the listener's auditory perception of movement between virtual sound sources from the first-person point of view in virtual reality applications. As a result, this paper serves to report only the additional work as performed in Part 2 of the study, though reference to Part 1 experiments will often occur. It is therefore suggested that the work presented in the publication on Part 1 be reviewed to gain full comprehension and background of the study's experimental procedures and data [125].

4.2 Introduction

Virtual reality technology provides a 360-degree, interactive immersive experience through the delivery of multisensory spatial awareness. While graphically modeled 3D environments can enhance spatial awareness and a sense of reality through the presentation of visual information, the same is true when sound is presented in 3D.

VAEs, which model the behavior of sound within virtual spaces, present spatial awareness through acoustically generated audible events [110]. These events, or spatial audio cues are related to the sound's interaction with the acoustic space. For example, early reflections help the listener to perceive the sound with distance and direction while reverberation adds a sense of overall room

dimension [41] [126]. In general, early reflections occur up to 80ms after the direct sound [41] [127]. As these reflections begin to regenerate and late reflections form more frequently with time, reverberation begins until the sound eventually decays to inaudible levels [128]. An illustration of this can be seen through the example of a room impulse response in Figure 23. While early reflections are known to increase localization cues [129], an extension of the early reflection period can have the inverse effect, especially with the addition of a strong reverberation period [130] [131].

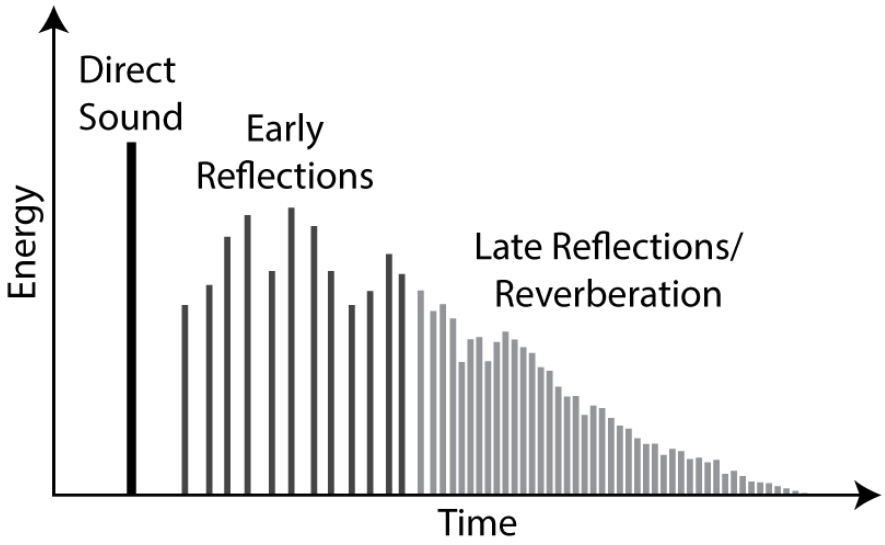

Figure 23: Diagram of a room impulse response

Modeling the complex behavior of sound to accurately generate these events for precision delivery of spatial audio cues is a significant challenge for developers. In fact, once motion is involved (giving the user the ability to navigate a virtual environment in real-time), performance and quality of the spatial audio becomes compromised. This is a problem directly related to the level of

complexity in the VAE and the constraints of the spatial audio processor. Moreover, poor rendering and simulation of listener movement within these VAEs can lead to audible artifacts (from miscues) in the portrayal of "realistic" 3D sound [28].

Two of the most important audible cues necessary for 3D localization and auditory motion perception are the ILD and ITD. These values are determined by the change in volume and time of arrival, respectively, from ear to ear for a given sound source located around the head [14]. Simply changing these values in real-time for binaural audio simulations often presents undesirable effects such as comb filtering and localization blur based on multi-positional, interpolated delays. Since sensing motion from audible cues is something we take for granted in real-world situations, when motion is produced incorrectly in virtual reality, these additive miscalculations of reality become immediately apparent.

Although an abundance of binaural audio research exists, most of this work was conducted to determine sound source localization [129] [132] [133] [134], basic sound object motion around the listener—from early physical experiments [49] to recent digital studies [120], and algorithmic modeling approaches for 3D sound reproduction [75] [135]. Additional work is required to specifically focus on the performance and quality of listener motion from these binaural simulation techniques. As such, this study intends to serve as foundational research in a new approach to understanding this topic. This multi-part study is investigating how auditory listener motion is perceived when moving between two static sound sources in a VAE. The basic sensation of motion perceived by the listener is the measure used to evaluate the performance of binaural crossfading

simulations against real-world binaural motion recordings, and motionless simulations.

4.3 Background

4.3.1 Auditory Motion Processes in Virtual Reality

Simple binaural crossfading is a technique that is commonly used in virtual reality applications to simulate the listener's perspective of auditory motion while moving from one virtual sound source to another. Simulating motion through audio can be achieved through a relatively simple process of crossfading between spatial audio cues, which are captured or simulated from two positional extents on a given path.

In practice, simple changes in reverberation and early reflections are often provided as sufficient simulations of listener motion between two acoustic positions. This technique is not 3-dimensional but is used often in game engines as a solution for motion perception and acoustic awareness. Inside the game engine, a reverb zone [136] is created as an area of graphical 3D space designed to generate specific acoustic parameters for additive reverb generation. When a listener enters a reverb zone, a function is triggered, which applies a customized reverb to all sound sources heard within the zone by the listener. At another point in space represented by a second reverb zone, the same audio will behave differently due to the changes presented by the acoustic parameters of the new reverb zone. As a listener passes from one zone to the next, the parameters are either crossfaded, or the output of the reverb is attenuated giving the listener the perception of movement from one VAE to the next.

Likewise, with binaural audio, panning the localization of sound objects in 3D space within a single VAE can be easily performed through crossfading the positional HRTF convolution processes [68]. This method is far superior to simply changing the reverberation characteristics but is still very basic in operation. Due to its simplicity, this technique is commonly used in virtual reality engines as an auditory motion solution. Since virtual reality content is quickly becoming a commercially accessible product, it is necessary to evaluate the performance and perception of these simple binaural crossfading processes. By understanding how simulations compare to the real-world, one can better determine the performance and/or pitfalls of this technology.

4.4 Method

For this multi-part study, an experiment was devised to both capture and simulate auditory listener movement from real-world physical motion in various acoustic environments. The goal was to evaluate basic binaural crossfading as a motion simulation process, and to analyze how reverberation affects the sensation of motion perceived by the listener. For all experiments in this multi-part study, the author maintained a consistent and repeatable experimental design, which controlled the methods used in the measurement procedure [125]. Therefore, the only change in Part 2 was that the measurements were conducted in a different acoustic environment, which is explained in further detail in Section 4.4.2.

4.4.1 Summary of the Replicated Measurements

The study accurately and consistently controlled a measurement process, which moved a motion

apparatus along a horizontal path, to capture the binaural auditory motion between two synchronized loudspeakers. This motion was captured in both directions (left to right, and right to left). The binaural recording was performed with a HATS mounted securely on the motion apparatus at a height of 1.5 m from the ground (loudspeakers were aligned to 1.5 m as well). The source audio signals used in Part 1 (stereo jazz, mono pink noise, and mono male speech) were reproduced through the loudspeakers as the stimuli for the binaural recordings. These binaural recordings became the experimental reference for all simulations.

HRTFs were gathered statically at the opposite extents of the motion path at listening position 1 and listening position 2 (see Figure 15). These measurements represented the localization cues for the two simultaneous sound source positions in space. Using motion-tracked video of the binaural motion recordings as a reference for position over time, auditory motion simulations were then created through binaural crossfading. These experimental simulations were synchronized in time and position to match the binaural reference recordings.

A static BRIR captured at position 3 was convolved with the stimuli and represented an audible anchor variable presenting no motion. Finally, first-person POV video was made for later use in the Audio/Video listening test. This captured a visual of the motion path as if perceived from the HATS' line-of-sight.

As in Part 1, angular localization positions were captured through BRIRs at five static positions along the motion path. These positions included 45 degrees left and right, 26.6 degrees left and

right, and 0 degrees center. The positions would be used later to convolve with stimuli for the localization training pre-test defined in Section 4.5.

Figure 24: Motion apparatus setup for Part 2

4.4.2 Acoustic Environment

Part 1 of this study was performed in a semi-anechoic chamber in order to reduce the influence of acoustic reflections on audible motion and localization cues. This acoustically controlled environment had a total room volume of 124 m^3 and a reverb time (T30) of 90 ms.

In contrast, measurements for Part 2 were conducted in a highly reflective, reverberant space. This

space, which we will term the reverberant environment, was an unfinished tracking room in the McGill University recording studios (see Figure 24). It was temporarily used for storage during final studio construction. The non-uniform design, along with scattered storage objects and hard finished walls made the space an ideal environment for unpredictable acoustic reflections, providing complete opposition to the controlled acoustic environment used in Part 1. The reverberant environment had a total room volume of 293 m^3, nearly twice the semi-anechoic environment, and a reverb time (T30) that was 10 times the duration at 900 ms. At its longest dimensions, the room was 8.5 m in length and 4.9 m in width. As shown in Figure 49, the apparatus was centered in the room such that P1 and P2 were each 1.5 m away from the side walls. The same 2 m^2 setup between the apparatus and loudspeakers was retained from the semi-anechoic experiment, however, the room widened at the front-left due to an entryway which caused Loudspeaker 1 to be further away from the left wall (2.24 m) than Loudspeaker 2 was from the right wall (1.5 m). For the front/back alignment, the speakers were placed 2 m from the front wall (to replicate the semi-anechoic setup). The apparatus was 4.5 m from the back wall and 2.9 m away from the storage objects placed in front of the back wall. A photograph of the room in Figure 24 shows the orientation in greater detail.

The final BRIRs were set to a window of 907 ms (using a rectangular window) and incorporated the onset pre-delay of 5.92 ms (2.03 m) at P1 and P2, and 6.50 ms ± 0.02 ms (2.23 m ± 0.02 m) at P3. This asymmetrical space had strong, equal energy early reflections at roughly 4 ms (floor reflection); 14 ms, 21 ms, 25 ms (side walls and ceiling); and 40 ms (back wall) following the

onset signal—a potential problem for localization. Greater detail of the 6 BRIRs obtained for this experiment are presented in Figure 43 through Figure 48.

4.4.3 POV Video Reference

The experiment also included first person *point-of-view* (POV) video for reference in perceptual tests. Video was included in the experiment to investigate whether fixed, visual guidance of the motion path would improve the sensation of motion. As visual information can be seen as a potential catalyst for change in perception, and since this experiment is studying motion perception through audio, the author chose to limit the viewing ability for testing purposes in order to control this additional variable as much as possible. It has been demonstrated often that reduced *field of view* (FOV) in virtual environments inhibits spatial awareness [137] [138] [139], so by controlling the POV video to only show the front of the room from the HATS perspective, the author could prevent extreme bias from 360 degree investigation of the virtual space, thus preventing the influence of visual information from dominating the auditory investigation altogether.

4.5 Listening Test

Through two versions of a listening test, 59 semi-experienced subjects evaluated the audible motion examples created during the measurement stage. These subjects also performed a localization training pre-test to familiarize themselves with the listening test environment and software, and to determine the total percentage of localization accuracy among subjects when presented with the experiment's 5 positional HRTFs. As in Part 1, a threshold of the total average

localization accuracy of the subjects was set to 90%. A value above this threshold would validate the ability of the experiment's HRTFs to represent accurate virtual positions along the motion path, and further validate any motion simulations generated from the use of these HRTFs. Lower values would indicate possible problems in the HRTFs gathered in the experiment.

Listening evaluations were performed in an acoustically isolated listening suite using headphones and an exact replication of the listening test setup from Part 1, including the same Lateral Listener Movement testing application. However, this test only used audio examples from the measurements taken during Part 2 of the experiment to focus on the influence of acoustic reflections.

4.5.1 Replicating the Testing Method

As in Part 1, subjects participated in a double blind, MUSHRA-style evaluation of the auditory motion examples [77]. Using this method based on a 100-point CQS, the subjects were asked to rate the sensation of motion perceived by the presentation of the stimuli through 3 randomized examples of auditory motion (reference, simulation and anchor). These groups of examples were presented over 12 separate trials comprising 6 left-to-right passes, and 6 right-to-left passes along the motion path. The binaural auditory motion recordings served as the hidden reference; the binaurally crossfaded auditory motion simulations served as the independent variable; and the static BRIR convolutions served as the anchor (presenting no motion). Each of the simulated motion passes were timed to video reference of the reference recording at 9 positions along the motion path. Each motion path was roughly 8 seconds in duration. The greatest variance in path

duration between a stimulus motion path (L2R and R2L) was 333 ms for the Speech stimulus. All others fell below 225 ms. Table 8 presents greater detail of the position-based time intervals used in each of the crossfading processes for this experiment.

Of the 59 listening test subjects, 30 were selected at random to take an Audio Only version, while the other 29 participated in an Audio/Video version containing additional first-person video of the motion path. Subjects were assigned assessor identification numbers of "A###" for the Audio Only test and "V###" for the Audio/Video test.

4.6 Results

4.6.1 Responses

In both test versions, the Shapiro-Wilk normality test showed that the sample data did not come from a normally distributed population for all conditions. Response data was therefore tested for a p value of $p < .05$ through a one-way analysis of variance on the distribution of data. All responses from the Audio Only and Audio/Video tests showed a significant difference between their pairwise comparisons of the signal by condition, $H(2) = 635.56$, $p < .05$; and $H(2) = 500.91$, $p < .05$ respectively.

When evaluating the data shown in Figure 25, one can see that for both versions of the test, the pairwise comparisons of responses by condition for all signals show that the simulation provided a greater sensation of motion overall to the listener while the anchor provided (by a large margin) the least sensation of motion, as one might expect.

Breaking down this data, one can see that when evaluating the responses by signal for individual conditions, it is clear that both the simulation and reference provided a similar sensation of motion when male speech was presented (see Figure 26). Interestingly, Figure 27 shows that the presentation of male speech through the binaural motion reference reduced the sensation of motion significantly in comparison to the reference results recorded in Part 1. Moreover, the simulation results for male speech appear to be unchanged from Part 1 to Part 2.

Again, in Figure 26, one can see how the presentation of music and male speech through the binaural crossfading simulation gave similar measures of sensation of motion, but also how pink noise greatly enhanced this measure above all other test conditions (reference and anchor).

4.6.2 Localization Accuracy

As in Part 1, there was only one localization training pre-test version. Therefore, all subjects' responses were pooled together to determine total localization accuracy for the entire group. Each of the five virtual positions was presented twice to each subject, giving 354 total trials per position over the entire subject group. The results of the localization training pre-test provided a mean accuracy of 90.5%, slightly above the experiment's threshold of 90%. Despite one less subject participating in Part 2 of the study than in Part 1, more localization errors occurred in Part 2 with the greatest amount of error coming from the Left Center virtual position. An illustration of these results can be seen in Table 4.

4.6.3 The Effect of POV Video on the Reference

The sensation of motion was significantly reduced when stimuli were presented through the binaural motion reference condition with the addition of the POV video. Though the results are still proportional to the Audio Only test results, the ceiling of these results was reduced by 15-20% for individual signals (see Figure 26). This result was expected as the viewing angle was fixed, and head-tracked movement or investigation of the virtual space was not made possible for this experiment.

4.7 Analysis

The results from Part 2 for both versions of the listening tests demonstrate proportionality among data. This proportionality was present between the two test versions in Part 1 as well. In fact, Figure 27 shows that the trend exists among all data from Part 1 and Part 2.

The data suggests that the sensation of motion is a measure that can be used to detect the perception of listener movement in a VAE. Additionally, it seems probable that a binaural crossfading simulation is an acceptable process for creating a perception of motion in its basic form (sensation only), which is at least equal to real-world motion as recorded through a dummy head. However, the data does not reflect what the measure of sensation of motion means in terms of directional accuracy, class of motion, or listener preference, but it certainly provides validation for further study in determining these factors, and to eventually understand what factors determine one's perception of audio quality in VAEs.

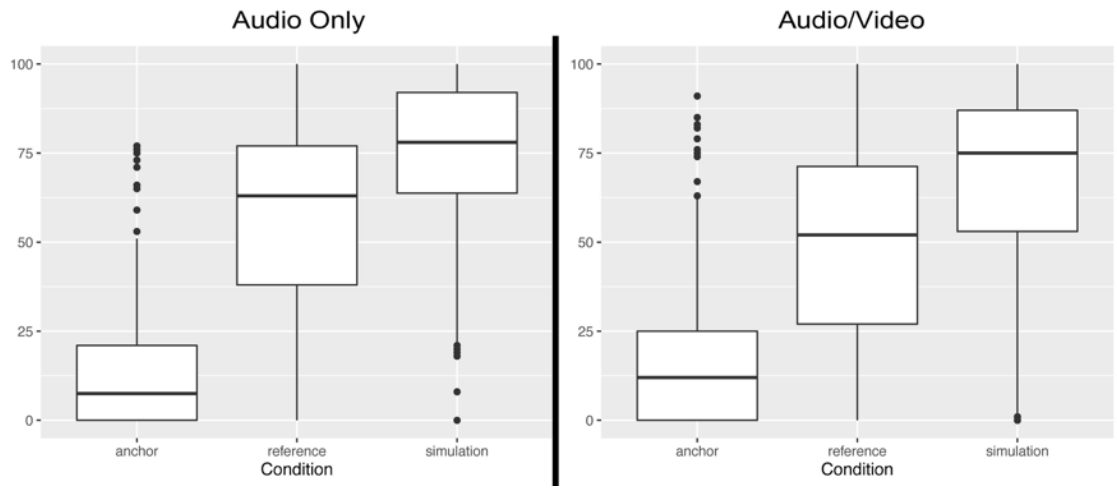

Figure 25: Responses by condition for all signals

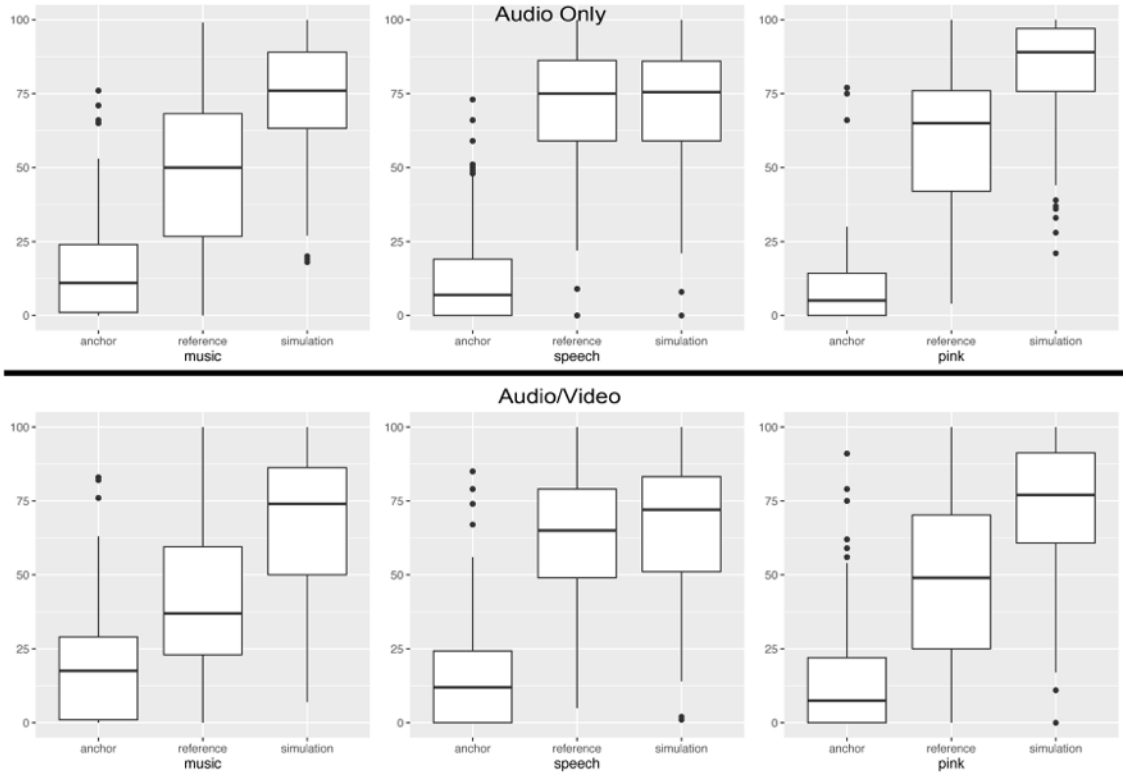

Figure 26: Responses by signal for individual conditions

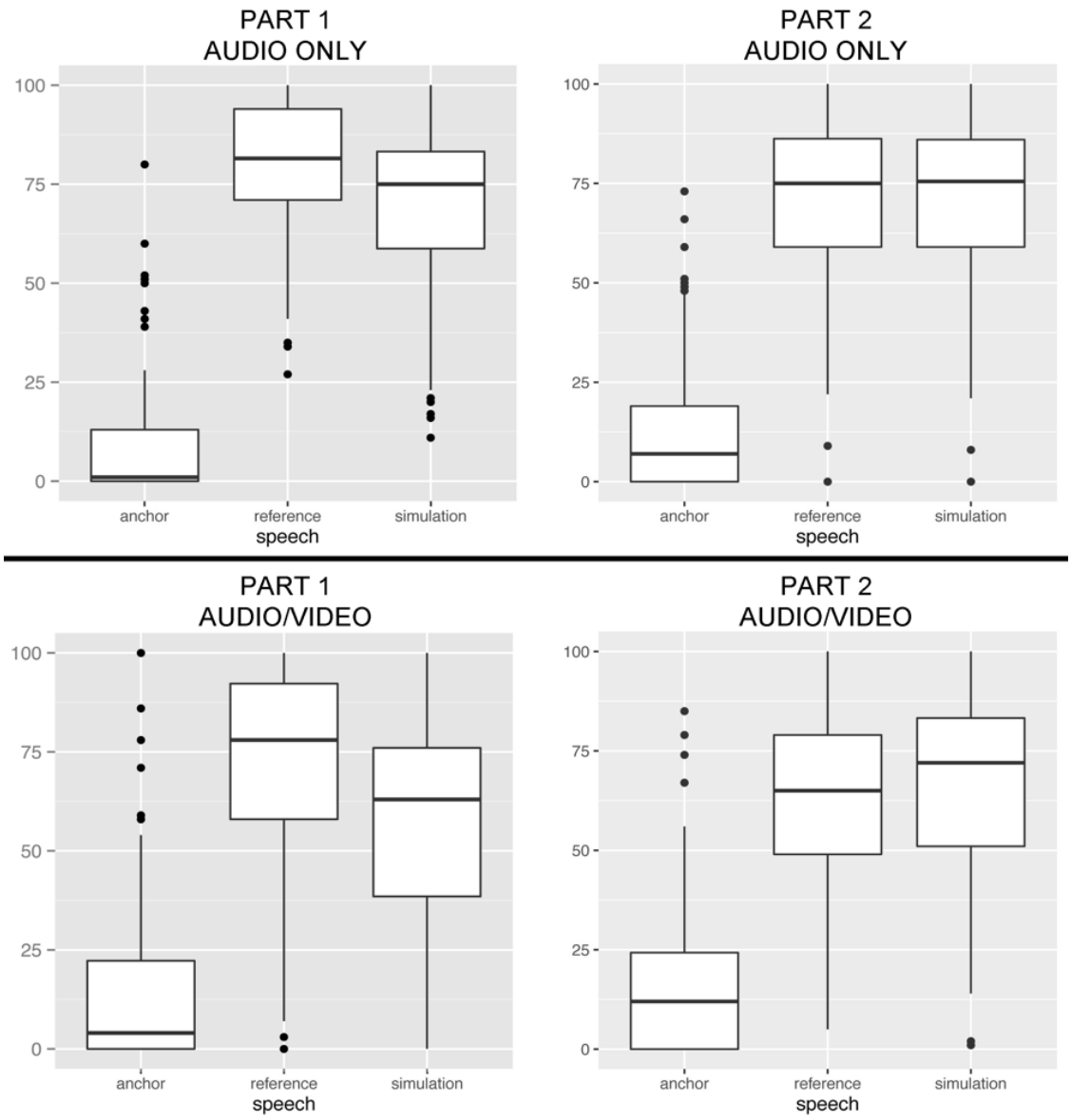

Figure 27: Comparison of Part 1 to Part 2 Speech responses for individual conditions

4.7.1 Reflections & Sensation of Motion

The presence of reflections seemed to improve the overall sensation of motion as presented by the

binaural auditory motion simulation. Reflections seemed to account for the increased measure of responses in the static anchor as opposed to responses in Part 1. It is interesting to note, however, that the reflections reduced the sensation of motion as produced by the binaural motion reference from the HATS recordings. Perhaps this is due to the exact delivery of the HRTFs from the HATS during the playback of these recordings versus the continuous averaging that occurs when these HRTFs are binaurally crossfaded for the simulation. It could then be proposed that binaural crossfading is a suitable method for overcoming the problems of non-individualized HRTFs for motion perception.

4.7.2 Video Influence

Reflections seemed to play a big part in the Audio/Video test as the presence of video in Part 2 seemed to confuse listening subjects on the apparent size of the virtual room (according to test comments in Section 4.7.4). The subjects felt that the reverberation did not match the real binaural recordings (reference), though this was naturally collected and reproduced in all examples. It is believed that the reduced visual awareness of the room in the video caused this confusion along with the influence of strong acoustic reflections which are known to cause localization blur. Further investigation of visual awareness will be performed in upcoming experiments.

Table 4: Localization accuracy for individual sound source positions, total localization accuracy (Part 2: Localization Pre-Test)

Sound Position	# of Trials	Accuracy	Errors	Total Accuracy
Left	354	87.6%	44	90.5%
L. Center	354	85.3%	52	
Center	354	90.1%	35	
R. Center	354	93.8%	22	
Right	354	95.5%	16	

4.7.3 Localization with Reflections

The localization accuracy of 90.5% for the entire subject pool satisfies the author as validation that the HRTFs were successful at providing accurate localization. However, it should be noted that the accuracy was reduced in comparison with Part 1's results of 92.7%. The long decay and high order, high energy reflections of the reverberant environment are assumed to have played a key role in this reduced accuracy by causing significant localization blur.

4.7.4 Listening Test Comments

The following comments are included in this paper as reference to the results and analysis:

> ***Assessor A120:*** *"B and C (reference & simulation) were equally good."*
>
> ***Assessor V105:*** *"It's distracting how the room from the visual doesn't match the visual's reverb."*
>
> ***Assessor V125:*** *"C (simulation) was disorienting, though it provided the most*

sensation of motion."

4.8 Future Work

With the successful completion of Part 1 and Part 2 of the multi-part study, the author suggests a follow-up experiment replicating the experimental measurement and procedure in virtual reality through an HMD. The design of the semi-anechoic chamber and reverberant environment in virtual space will allow for listening tests to be performed with precision to exactly match the scale of the setups in the previous experiments. This experiment will also allow for the listening subjects to visually investigate the entire measurement environment, unlike the presented videos in the Audio/Video trials of Part 1 and 2. The author hopes that future experiments will provide information on the importance of visual awareness when determining sensation of motion in a VAE and virtual reality environments. In addition, studies are suggested on directional accuracy in binaural crossfading simulations, as well as listener preference for specific techniques.

4.9 Conclusions

In Part 1 and Part 2 of this study, the simulation of auditory listener motion perceived by the participants consistently resulted in a measure of motion sensation that, on average, was rated similarly or greater than the results of the real-world reference. The presence of reflections in the experiment increased this measure for the simulation while reducing it for the real-world reference, and localization accuracy. Early acoustic reflections in high order, and high energy combined with the significant reverberation period seemed to have been a catalyst for this reduction of quality.

No sufficient evidence was found that would reject the null hypothesis of equivalence between the simulation and the reference.

The data suggests that the simulation was less influenced by reflections than the reference. However, one must not rule out the possibility that due to the simplified HRTF averaging that occurs during binaural crossfading, the greater sensation of motion in the simulation may in large part be due to a different motion transition experience than the reference. The comments from assessors (V125 & A120) seem to argue against a purely consistent experience that may be expected by linear crossfading, instead referring to a motion experience that is "good" but "disorienting" (see Section 4.7.4). Interestingly, the presentation of Male Speech through the reference exhibited equal or lesser measures of sensation of motion than the simulation, while in Part 1, sensation of motion by Male Speech was rated higher than the simulation. This data suggests that the sensation of motion results from the reference Male Speech were influenced by the addition of strong reflections. Further influence in this area may have come from the presentation of video with limited field of view. For example, the comment from Assessor V105 (in response to the Male Speech signal presented through the reference) complains that the "reverb" conflicts with the assessor's (limited) visual awareness of the space, thus causing distraction. This gives reason to consider that reference results might be influenced by a level of annoyance due to the limited FOV.

As a result, the third and final part of this research will investigate how visual awareness of virtual rooms through an HMD impacts the listener's sensation of motion.

5 The Effect of Virtual Environments on Localization during a 3D Audio Production Task

Abstract

In a perceptual study of three-dimensional audio for VR, an experiment was conducted to investigate the influence of virtual environments on localization during a production-based panning task. It was hypothesized that performing this task in VR would take longer and be less accurate than if conducted in a real environment. Using a 3D loudspeaker array and hardware panning controls, 80 participants were asked to repeatedly match probe sounds to the location of a randomly positioned target sound. Participants performed the task with and without awareness of loudspeaker position. Half were presented a VR replica of the environment using a head-mounted display. Results showed that virtual reality did not significantly inhibit task performance.

5.1 Introduction

The field of audio engineering is pivoting into the new medium of spatial audio (a presentation of

immersive sound in 360 degrees) [140]. Audio engineers are quickly responding to a commercial demand for spatial audio productions for virtual reality products. Though engineers can instantaneously audition real-time processes in recording studio productions, auditioning in virtual reality is not always real-time. Spatial audio is commonly produced outside of virtual reality. Before the audio production can be properly auditioned in 3D, the production must be compiled through a game engine along with video and programming code to create the final virtual reality product. To bypass this issue, engineers audition their spatial audio productions through middleware [141], giving a two-dimensional (2D) representation of the final product by emulating 3D control and interaction. However, much like a movie soundtrack produced without the re-recording process (final mix to video), spatial audio production for virtual reality can become guesswork. Since the engineer must create and compile before experiencing his changes, it becomes incredibly difficult to predict the appropriateness and synchronization of audio for the final 3D interactive product.

However, virtual reality technology is currently advancing audio production workflow. These advancements allow for audio production tasks to be performed entirely within virtual reality, and most importantly, while wearing an HMD, providing real-time auditioning. One such example for real-time visual control within virtual reality is the Tilt Brush [142], which lets users create real-time visual art in virtual reality by operating a peripheral hardware controller. This same idea can be applied to audio, essentially allowing for the same potential and giving peripheral control of interactive audio programming within the virtual reality application. Such interactive control, if

programmed in advanced, can allow for sound emitters to be placed or altered in a virtual environment in real-time, by the engineer. As use of these controllers for audio production become more conventional, and processing speeds increase for real-time audio operation, audio engineers with less access to professional equipment may even choose to control and produce audio entirely in virtual environments, leaving the real studio behind for easily accessed, virtual replicas [143].

5.2 Inspiration

In the field of audio production for film (commercial movies), it is not uncommon for sound design, recording and mixing stages to occur before visual productions are finalized. Audio engineers may need to depend solely on rough concept videos, storyboards, or sometimes without visuals at all. For this reason, most video productions include a "re-recording" stage to finalize the audio mix once the video or film has reached final production. Re-recording is typically performed on a sound stage (a replica movie theater) as this is the platform where the final production will be experienced after commercial release. The audio production, along with simulations of auditory motion provided by crossfading, is checked for final performance against the reference visual, ensuring accuracy and success in the overall audiovisual experience. However, in virtual reality, these situations do not apply.

It is uncertain how one might ensure accuracy in an audio production task (i.e., motion simulation) if conducted entirely within virtual reality, especially following commercial release where the production will be experienced by infinitely different and unique sets of ears via headphones, rather than through an averaged acoustic environment with loudspeakers, like a movie theater.

Furthermore, replication of the final production environment may be limited to device performance of different HMDs.

For spatial audio production created entirely within virtual reality, such variety of visual awareness is common due to the lack of available resolution, limitations of head tracking, and/or limitations in real-time performance in the virtual reality HMD. It is necessary for an audio engineer to be aware of such limitations that may be present during virtual reality audio production. A scientific analysis of task-related performance will help define better workflows and future standards in this newly emerging field.

5.2.1 Limitations of Reduced Visual Awareness and Linear Crossfading

Results of the experiments in Chapter 3 and Chapter 4 suggest that under certain conditions, the sensation of motion in the binaural conditions may have been influenced by a lack of visual awareness. So, it would follow that the lack of visual reference in virtual reality might negatively influence audio production tasks based solely within virtual reality. In the previous experiments, simulated auditory motion produced through binaural crossfading reported less visual influence on the sensation of motion than the referenced real-world binaural auditory motion. In addition, little evidence was found to suggest binaural crossfading provides a less than equivalent sensation of motion to the real-world binaural reference. These results give reason to further the study of the crossfading method in Chapter 3 and Chapter 4 for simulating auditory motion in binaural processes. To advance this study, one might focus on the technique of crossfading as an audio production "task" to achieve motion sensation and develop a greater understanding of the influence

of visual awareness on this task in virtual reality applications. Since linear crossfading was suspected of causing inconsistent motion accuracy in previous experiments due to the plausibility of perceptual "jumps" from one localized position to another, a more accurate method of crossfading is hoped to provide greater assurance in the experimental results.

5.2.2 Audio Source Panning

In audio production, the task of panning a point source from one loudspeaker to another is as routine as asking a musician to play a note on their instrument. Positioning a sound source in a sound field creates separation, depth and can often be used as a technique to present movement cues to the listener. When performing this audio task for video (i.e., post-production projects), engineers typically pan important program material to match the location of visual objects emitting sound on screen. This technique has been widely adopted since stereo sound fields pair well with standard, 2D video productions [144]. However, virtual reality requires sound localization to match visual objects in 3D space, which is a 3D problem that only spatial audio can solve [145] [32].

5.2.3 Technical Uncertainties

Stereo techniques only provide audio source panning up to 90 degrees left or right of the listener. Panning in 3D presents the ability to position sound objects at 360 degrees around the listener, presenting a much larger opportunity for error. Studies reported by Blauert show visual cues shift auditory localization towards the visual emitter; however, these studies were most often performed with real-world visuals or non-immersive video [146] [88]. In virtual reality, FOV is often limited

and definition of video is significantly blurred, pixelated or distorted in comparison to its counterpart in reality. The visual presentation of the virtual environment alone, in this less-than-ideal scenario, may have an effect on auditory localization [147]. In addition, the larger positional field may also increase difficulty or visual-cue shift in localization accuracy.

Furthermore, 3D panning techniques require more angular precision of amplitude distribution between loudspeakers than the linear and logarithmic panning techniques used in stereo. As a result, several methods exist for source panning in 3D, including basic *linear panning* (LP), VBAP, *multiple-direction amplitude panning* (MDAP) and Ambisonics gain-factor panning, as reported in [148] [149]. Hardware and software controllers can be used to operate source panning in 3D with these techniques, but task performance may be limited due to inaccuracies in scale between the control system and the virtual environment's audio emitters (or loudspeakers).

Previous research has focused on new techniques for quality and performance. Little work has been done to report the influence of these uncertainties on virtual reality audio production.

5.3 Method

It was assumed that virtual reality and its technical shortcomings would negatively affect the production workflow of an audio engineer. A study was conducted to test the hypothesis that localization accuracy and task performance would be reduced when listeners perform a 3D audio panning task in virtual reality.

5.3.1 Experimental Listening Environment

An experimental listening environment (see Figure 28) was created in a semi-anechoic chamber by constructing a hemispherical 3D loudspeaker array with a central listening position. The array was based off the NHK 22.2 standard for multi-channel sound [99] and contained 17 Genelec 8030 loudspeakers around a central listener position. All loudspeakers were calibrated to a listening level of 65 dB SPL (A-weighted). The loudspeakers were positioned 45 degrees apart and separated over 3 vertical layers. The bottom layer was at ear-level (0ºN) and contained 8 loudspeakers. The top layer was positioned 38ºN above the bottom layer and contained another 8 loudspeakers. The overhead layer contained one loudspeaker positioned directly above the listener, at 90º N, perpendicular to the bottom layer. Loudspeakers in the first two layers were positioned at a horizontal distance of 1.91m from the listener position. The bottom layer was positioned at a height from the floor of 1.22m, with the top layer at 2.72m and the overhead layer at 3.05m. All loudspeakers were active during the experiment. Figure 29 shows an example of a testing participant seated at the listening position in the experimental environment.

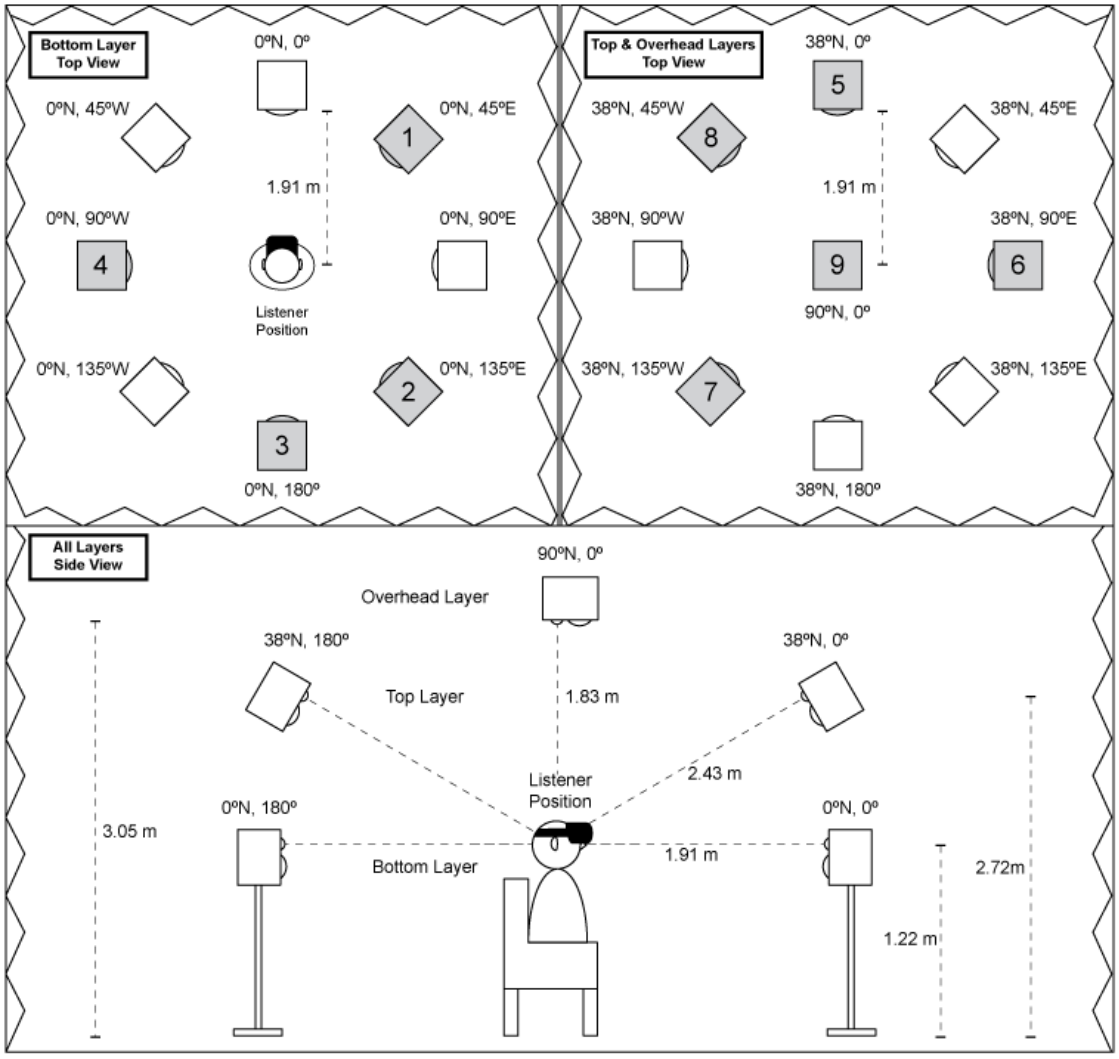

Figure 28: Diagram of the experimental environment for sound source panning and localization

5.3.2 System for 3D Source Panning

3D audio panning operation was provided through the IRCAM-SPAT 3D panning software [150], using the VBAP method [61]. The panning software was controlled by two Griffin PowerMate

USB hardware rotary controls. These controls offered continuous rotation without notched or physical reference of default position. One control was used for elevation panning and another for azimuth. Audio sources could be panned at intervals of 5 degrees in azimuth and intervals of 1 degree in elevation.

Figure 29: Panoramic view of actual experimental environment

5.3.3 Virtual Environment Design

A professional 3D modelling/texture artist designed a virtual replica of the experimental listening environment (see Figure 30) and its objects (wall material, fixtures, and loudspeakers) from specified material lists, photos and architectural measurements. The outcome of the virtual environment was that it reflected a true, high-resolution virtual copy of the real environment. Using an Oculus Rift HMD and Unity 3D software, the virtual environment was presented in virtual reality at the listening position. Location of loudspeakers and room dimensions were calibrated to

match the FOV, depth and position of the real environment. In the virtual environment, the room and its objects were built to scale and compared to architectural and physical measurements of the real room, using precision measurement devices. Room extents and loudspeakers were aligned visually as a final calibration. Regular users of the actual room were also consulted and confirmed the virtual reality replica presented precision realism. This created a one-to-one match of experimental variables in real and virtual environments.

5.3.4 Experimental Conditions

The experiment presented 4 listening conditions. Loudspeaker objects in the virtual environment could be hidden to present a virtual environment without the presence of visual sound objects. This setup represented the “virtual—blind” (Vb) condition. With the virtual loudspeakers present, the “virtual—speakers” (Vs) condition was presented (see Figure 31). Listeners in the real room, without virtual reality HMDs saw the actual loudspeakers and the real environment. This represented the “real—speakers” (Rs) condition. Some listeners in the real room were blindfolded. These listeners were neither aware of the loudspeakers, nor the environment. This represented the “real—blind” (Rb) condition.

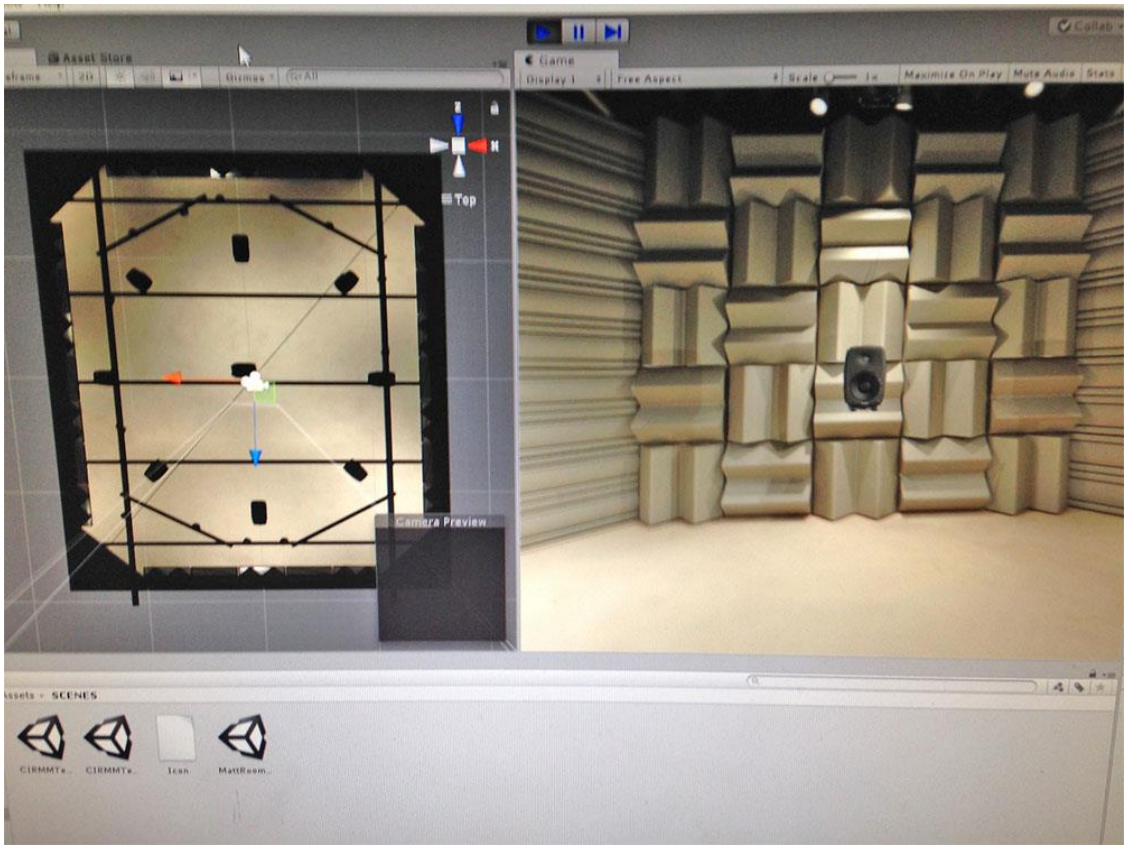

Figure 30: A 2D screenshot image of the game engine which presented the 3D virtual environment to the listening subjects in virtual reality via the HMD

5.4 Experiment

In experimental trials, 80 trained participants were asked to perform a 3D audio panning task to match the location of a probe sound to the location of a target sound. This task was replicated over four separate conditions and recorded localization error and task duration for each task.

5.4.1 Stimuli

Probe sounds were loudness matched to the target sound [115]. No visual feedback cues were

given as to the current location of the probe sound or target sound; participants performed the task based on localization cues, alone. The probe sound consisted of two randomly presented audio loops (female voice and conga drums). Short pink noise bursts were presented as the target sound. These bursts were based on a previous localization methodology to reduce perceptual drift [151].

5.4.2 Participants

The participants ranged in age from 18 to 34, all had at least one-half year of audio engineering education (most with a background in music performance), and all with basic technical ear-training experience. Participants were chosen at this skill level for their relatively novice level of spatial audio production for VR to reduce prior experience biases to 3D audio systems and VR. All participants reported normal hearing. Several participants had natural vision impairments but used glasses or contacts to regain full vision.

At random, participants were grouped into one of the four experimental conditions (Vs, Vb, Rs, and Rb). Participants did not perform multiple conditions and had no previous knowledge of the difference between conditions, or of the experimental environment. Participants were first trained on practice trials to learn the controls and to become comfortable with environmental conditions.

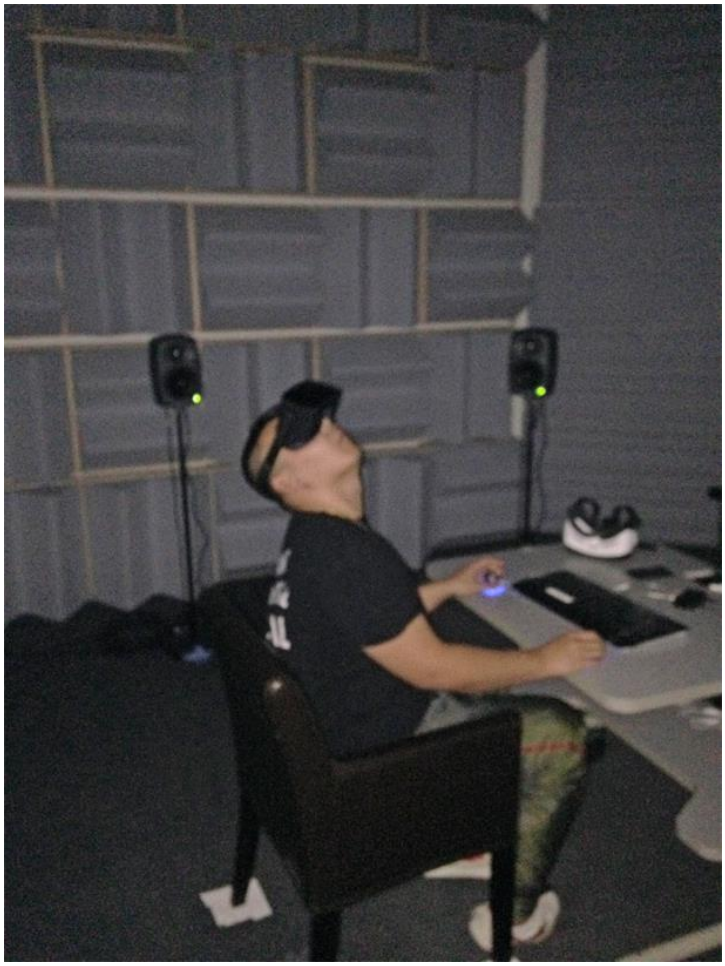

Figure 31: Vs participant experiencing the virtual environment with virtual speakers while listening to real speakers

5.4.3 Localization Task

For each trial, a target sound was randomly presented at 1 of 9 fixed loudspeaker locations in the array. For reference, darkened speakers with corresponding numbers and angular positions in Figure 28 denote these locations. Immediately following the presentation of the target, a randomly selected probe stimulus (voice/conga) was presented directly in front of the participant at 0ºN, 0º.

Using the rotary panning controls, the participant moved the probe sound to match the localization of the probe to the target. Participants were allowed to move their head freely. Once the participant thought the two sounds were localized together, a localization answer was submitted, and a new trial began. This process was repeated 5 times for each of the 2 probe sounds, at each of the 9 target positions, yielding 90 total trials per participant. The experiment lasted on average 45 minutes. Participants were encouraged to take breaks.

5.4.4 Recorded Data

For each participant, task performance was recorded through localization accuracy (angular error) and response time (task duration in ms) across each of the four experimental conditions at each of the 9 locations. Latitudinal (elevation) error and longitudinal (azimuth) error were recorded separately for analysis of directional accuracy. These values were used to calculate absolute angular error (calculated as the central angle between two coordinates on the great circle) using Equation 3 [152] where θ is the central angle, $\emptyset_1$ and $\emptyset_2$ are the latitude angles of the two points, and λ_1 and λ_2 are the longitude of the two points. Unless otherwise stated, error data in this chapter is reported in absolute values.

Equation 3: Formula for the central angle on a great circle

$$\theta = \arccos(\sin\emptyset_1 \cdot \sin\emptyset_2 + \cos\emptyset_1 \cdot \cos\emptyset_2 \cdot \cos(|\lambda_1 - \lambda_2|))$$

5.5 Results

Dependent variables (accuracy and duration) were analyzed for normality within the separate

experimental conditions (Vs, Vb, Rs, Rb) and the grouped conditions (All Blind, All Speakers, All Real, All Virtual). In addition, accuracy results were analyzed for normality at each location. The population of the accuracy and duration data in each of the said conditions did not fit a normal distribution as tested with histogram analysis and the one-sample Kolmogorov-Smirnov non-parametric normality test [153], at the 95% confidence interval. As the results were of non-normal distributions, an analysis of medians was conducted, rather than means [154].

5.5.1 Noise Reduction

Before performing normality and significance tests, a method for noise reduction was considered for the recorded data. This method was used to determine whether it was necessary to remove extreme outliers from the dataset. Data extremes related to improbable responses (user and machine error for duration) and the inability to localize (beyond 90 degrees of error) were evaluated. For individual results within each of the 4 experimental conditions, Equation 4 [155] was used to calculate extremes, where IQR is the calculated *interquartile range* and k = 3 is the multiplier used to indicate extreme outliers at values (i) in the data array (X).

Equation 4: Formula for the calculation of data extremes using the interquartile range

$$X_i = [Q_1 - k(\mathrm{IQR}), Q_3 + k(\mathrm{IQR})]$$

Extremes were insignificant in the calculation of medians and means at the 95% confidence interval; testing against values outside this range will lead to p-values < 0.05. For the purpose of visualizing data falling within the 95% confidence interval, outliers and extremes were removed

from all statistical figures included in this chapter.

5.5.2 Significance Tests

Significance in pairwise comparison of the test results was calculated using either the Kruskal-Wallis comparison of medians, or the Mann-Whitney U-test, when appropriate [154]. An alpha level of .05 was used for all statistical tests.

5.5.3 Results by Probe Stimuli

A comparison of medians for accuracy (error in degrees) and duration (seconds) were made between probe 1 (female voice) and probe 2 (conga drum) results. The presentation of female voice resulted in slightly more localization error (11º), $H(2) = 5.69$, $p = .017$, than conga drums (10.62º). Task duration was longer for female voice trials (14.09s), $H(2) = 18.19$, $p < .001$, as compared with conga drum trials (13.20s).

5.5.4 Results by Location

An analysis of results by location was performed to understand any possible location-based influences existing between real and virtual conditions (see Table 5 and

Table 6). Figure 32 details the range of error results across all conditions for localization accuracy at each location; plots are grouped by similar locations to aid visual comparisons.

5.5.4.1 Accuracy by Location

Location-based results for localization accuracy were compared across real and virtual conditions, blind conditions and speaker conditions. Results for Location 1 ($H(2) = 4.48$, $p = .03$) and Location 5 ($H(2) = 8.05$, $p = .005$) each varied by 2 degrees of localization error between real and virtual conditions. For blind conditions, Location 2 ($H(2) = 8.00$, $p = .005$), Location 5 ($H(2) = 6.14$, $p = .01$), and Location 8 ($H(2) = 4.97$, $p = .03$), varied by 4.1, 2.66, and 2.79 degrees, respectively. For speaker conditions, Location 3 ($H(2) = 4.87$, $p = .03$) was the only location with a significant variance in localization accuracy results, with an error that varied by 1 degree.

5.5.4.2 Duration by Location

Duration results differed significantly at each of the 9 locations when compared across real and virtual conditions, $H(2) = 4.41 \sim 21.61$, $p \leqq .04$. This was also true at all locations for the blind condition comparisons, $H(2) = 13.85 \sim 39.40$, $p < .001$. However, there was no significant difference in results among speaker conditions for any of the locations, $H(2) = .04 \sim 1.64$, $p \geq .20$. These results have been summarized for brevity and clarity within this paragraph; however, individual location-based results and their exact p-values can be viewed in greater detail in Table 5 as well as the results summarized across condition groups in

Table 6.

5.5.5 Results by Condition

A final grouping of the results by condition was performed to understand the total influence (if any) of the experimental conditions on task performance results.

5.5.5.1 Accuracy by Condition

Localization accuracy was evaluated by comparing localization error results across each of the grouped conditions. This analysis showed a slight significance in the variance of localization error for real and virtual conditions, and within blind conditions. Accuracy results between real and virtual conditions differed by .53 degrees ($H(2) = 6.01$, $p = .01$) with the median error at 11º for real conditions and 10.47º at virtual conditions.

When comparing the results of the blind conditions (Rb vs. Vb), localization accuracy differed by 1.36 degrees ($H(2) = 6.46$, $p = .01$) with the median error at 12.36º for the real blind condition and 11º for the virtual blind condition.

Speaker conditions did not reveal a significant difference in localization accuracy results ($H(2) = .81$, $p = .37$) and differed only by .29 degrees with the median error at 10º for the real speaker condition and 9.71º for the virtual speaker condition.

5.5.5.2 Duration by Condition

Task duration was evaluated by comparing results across each of the grouped conditions. This analysis showed significance in the variance of duration for real and virtual conditions, and within

blind conditions. Duration results between real and virtual conditions differed by nearly 2 seconds (1.998s) ($H(2) = 121.98$, $p < .001$) with the median duration at 14.848 seconds for real conditions and 12.85 seconds at virtual conditions.

Task duration for blind conditions differed by 4.202 seconds ($H(2) = 188.17$, $p < .001$) with the median duration at 18.285 seconds for the real blind condition and 14.083 seconds for the virtual blind condition.

Differences in task duration results ($H(2) = 3.85$, $p = .0497$) differed only by 260 ms with the median duration at 11.715 seconds for the real speaker condition and 11.455 seconds for the virtual speaker condition. It should be noted that the difference in medians is technically significant based on a p-value that has been expressed to 4 decimal points. The probability that the difference in medians is outside the 95% confidence interval is based on a value of 3 ten-thousandths (.0003). Rather than speculate or attempt to skew results that don't seem to follow the significance reported in the location-based results of Section 5.5.4.2, this p-value has been rounded up to $p = .05$, and thus it is considered that the medians of task duration results for speaker conditions exhibit no significant difference.

Table 5: Median values for accuracy and duration by location for all condition groups

Loc.	Elev.	Azim.	Median Error (Real)	Median Error (VR)	H(2)	p < .05	Loc.	Elev.	Azim.	Median Duration (Real)	Median Duration (VR)	H(2)	p < .05
Accuracy (°) By Location for Real vs. Virtual Conditions (R/V)							**Duration (ms) By Location for Real vs. Virtual Conditions (R/V)**						
1	0°	45°E	*9.00*	*7.00*	*4.48*	*p = .03*	1	0°	45°E	*10950.0*	*9052.5*	*21.61*	*p < .001*
2	0°	135°E	11.00	10.14	1.56	p = .21	2	0°	135°E	*14705.0*	*12042.5*	*16.54*	*p < .001*
3	0°	180°	8.00	8.00	2.39	p = .12	3	0°	180°	*15357.5*	*14212.5*	*4.41*	*p = .04*
4	0°	90°W	10.90	10.72	0.27	p = .61	4	0°	90°W	*11190.0*	*9370.0*	*11.56*	*p < .001*
5	*38°N*	*0°*	*13.00*	*10.96*	*8.05*	*p = .005*	*5*	*38°N*	*0°*	*14812.5*	*12077.5*	*19.13*	*p < .001*
6	38°N	90°E	13.39	13.16	0.90	p = .34	6	38°N	90°E	*17990.0*	*15567.5*	*13.34*	*p < .001*
7	38°N	135°W	15.53	16.00	0.52	p = .47	7	38°N	135°W	*19545.0*	*17140.0*	*12.80*	*p < .001*
8	38°N	45°W	13.07	11.69	2.94	p = .09	8	38°N	45°W	*15817.5*	*13410.0*	*19.55*	*p < .001*
9	90°N	0°	0.00	1.00	1.82	p = .18	9	90°N	0°	*14157.5*	*11592.5*	*19.63*	*p < .001*
Accuracy (°) By Location for Blind Conditions (Rb/Vb)							**Duration (ms) By Location for Blind Conditions (Rb/Vb)**						
1	0°	45°E	9.92	7.81	2.03	p = .15	1	0°	45°E	*14355.0*	*10737.5*	*39.40*	*p < .001*
2	0°	135°E	*15.10*	*11.00*	*8.00*	*p = .005*	2	0°	135°E	*18722.5*	*14700.0*	*15.62*	*p < .001*
3	0°	180°	8.97	8.97	0.00	p = .95	3	0°	180°	*18805.0*	*15577.5*	*16.05*	*p < .001*
4	0°	90°W	11.70	11.32	0.47	p = .49	4	0°	90°W	*14642.5*	*10607.5*	*22.49*	*p < .001*
5	38°N	0°	*13.53*	*10.87*	*6.14*	*p = .01*	5	38°N	0°	*17552.5*	*13102.5*	*34.06*	*p < .001*
6	38°N	90°E	13.02	13.90	0.27	p = .61	6	38°N	90°E	*22485.0*	*16897.5*	*21.45*	*p < .001*
7	38°N	135°W	17.00	15.87	0.20	p = .66	7	38°N	135°W	*22997.5*	*19157.5*	*13.85*	*p < .001*
8	38°N	45°W	*14.75*	*11.96*	*4.97*	*p = .03*	8	38°N	45°W	*20972.5*	*14875.0*	*28.44*	*p < .001*
9	90°N	0°	2.00	5.00	0.74	p = .39	9	90°N	0°	*16532.5*	*12192.5*	*24.49*	*p < .001*
Accuracy (°) By Location for Speaker Conditions (Rs/Vs)							**Duration (ms) By Location for Speaker Conditions (Rs/Vs)**						
1	0°	45°E	7.00	7.00	2.24	p = .13	1	0°	45°E	7777.5	7482.5	1.22	p = .27
2	0°	135°E	8.00	9.50	1.59	p = .21	2	0°	135°E	11090.0	10720.0	1.49	p = .22
3	0°	180°	*8.00*	*7.00*	*4.87*	*p = .03*	3	0°	180°	12582.5	13205.0	1.26	p = .26
4	0°	90°W	9.00	9.10	0.00	p = .99	4	0°	90°W	8362.5	8492.5	0.04	p = .84
5	38°N	0°	12.26	10.96	2.23	p = .14	5	38°N	0°	11345.0	11105.0	0.84	p = .36
6	38°N	90°E	14.07	12.59	3.71	p = .05	6	38°N	90°E	14487.5	14072.5	0.32	p = .57
7	38°N	135°W	14.62	16.28	2.22	p = .14	7	38°N	135°W	17590.0	15785.0	1.64	p = .20
8	38°N	45°W	12.12	11.46	0.02	p = .88	8	38°N	45°W	12247.5	11745.0	1.18	p = .28
9	90°N	0°	0.00	0.00	1.32	p = .25	9	90°N	0°	11545.0	11072.5	1.55	p = .21

Table 6: Median values for accuracy and duration across condition groups

Condition	Median Error		H(2)	p < .05	Median Duration		H(2)	p < .05
	Real	VR			Real	VR		
	Accuracy (º) by Condition				Duration (s) by Conditions			
Real vs. Virtual	11.00	10.47	6.01	p =.01	14.848	12.850	121.98	p <.001
Blind	12.36	11.00	6.46	p =.01	18.285	14.083	188.17	p <.001
Speakers	10.00	9.71	0.81	p =.37	11.715	11.455	3.85	p =.0497

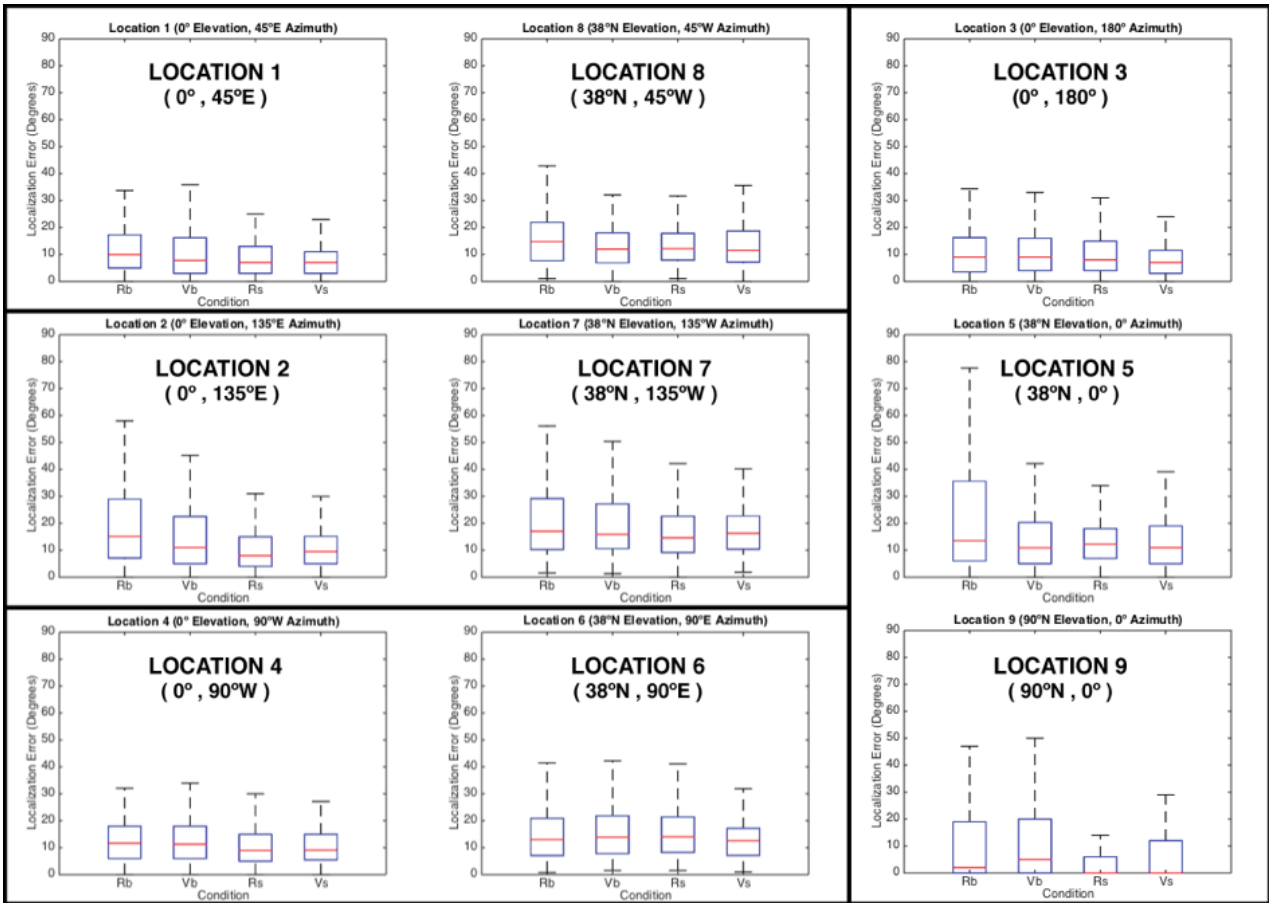

Figure 32: Distribution of individual localization accuracy results by location

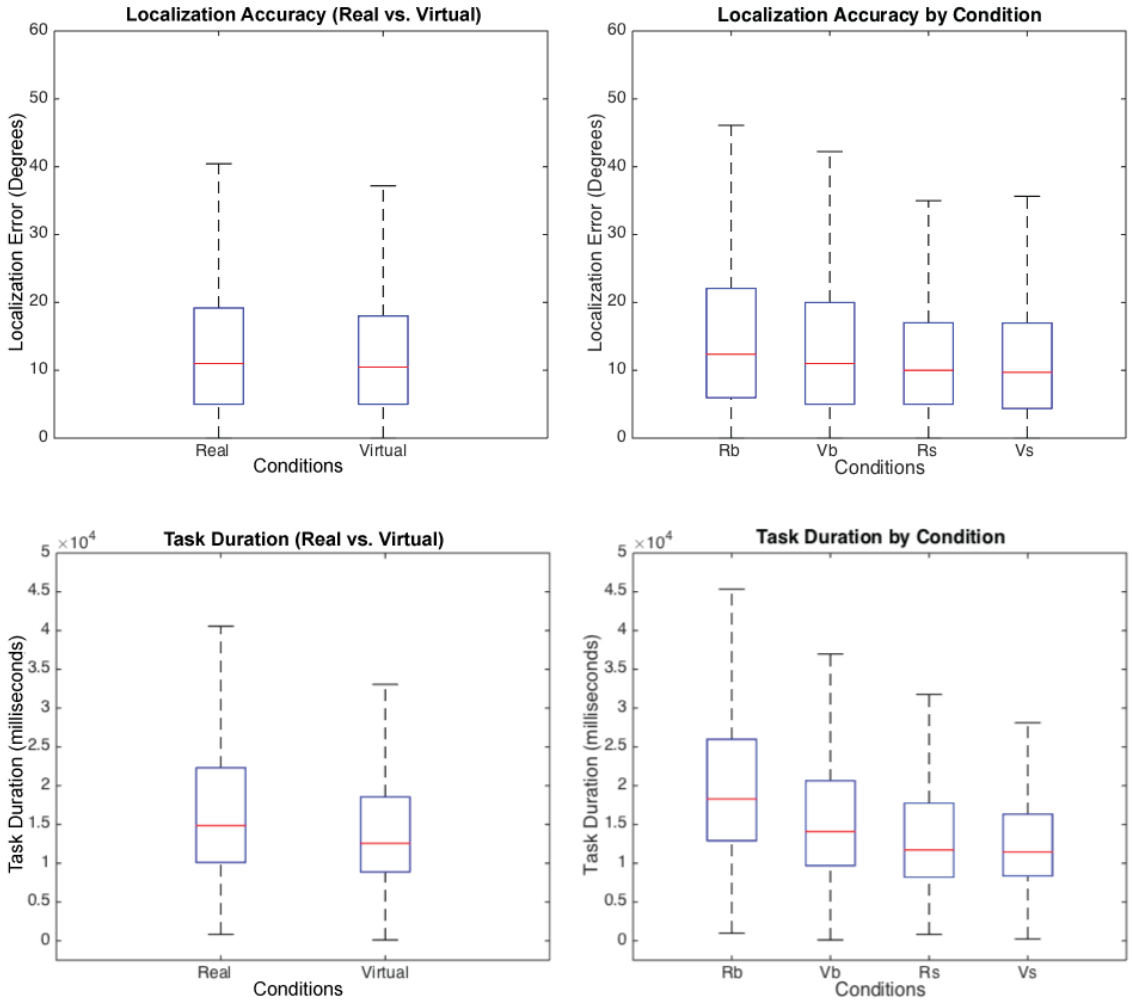

Figure 33: Distribution of localization accuracy (localization error in degrees) and task duration (ms) results by condition. On top, from left to right: localization accuracy (real vs. virtual), localization accuracy by all conditions (Rb, Vb, Rs, Vs). On bottom, from left to right: task duration (real vs. virtual), task duration by all conditions

5.5.6 Effect of Duration on Localization Accuracy

A simple linear regression was calculated to predict localization accuracy (error in degrees) based

on task duration (seconds). A significant regression equation was found ($F(1,7198) = 13.857$, $p < .001$), with an R2 of .002. Participants' predicted accuracy is equal to 16.678º - .068º (duration) when duration is measured in seconds. Participants' accuracy improved (error decreased) by .068º for each second of task duration.

5.6 Discussion & Future Work

When the participant could not see at all (Rb), results were largely affected by larger error and longer duration, as shown in Figure 33. This agrees with results from previous studies on visually-aided localization task performance [106] [107]. For example, awareness of the virtual environment, even without visual emitter reference (Vb), significantly improved task performance over Rb. The results also seem to suggest that the significant difference that exists between grouped real and virtual conditions for localization accuracy and task duration is largely due to the variance of the skewed Rb condition results. Finally, speaker conditions showed no significant difference.

Considering the results of these specific tests, it can be concluded that in this experiment, by creating a virtual replica of a real environment, the participants of the 3D panning task were not impeded by virtual reality in their task performance. Therefore, the hypothesis was false. This experiment suggests that if virtual environments are created to replicate reality, an audio engineer working in virtual reality might expect similar, unobstructed performance in a localization production task.

Due to length restrictions for this publication, the large amount of data gathered in this experiment

was reduced in reporting to focus specifically on environmental influence on real vs. virtual condition results. Therefore, the directional tendency of accuracy results around the target location cannot be concluded, however a preliminary analysis has shown possible significance in an upward shift in elevation-based results between real and virtual conditions. Azimuth results had a tendency to appear slightly left of the target. Analysis of participant experience as it relates to results was not reported and could be helpful in further noise reduction and statistical analysis.

Bias from end-point effects [113], due to the constraints of the 3D panning system and the loudspeaker layout should be noted to have occurred in this experiment. This was present due to the fact that our loudspeaker array did not contain loudspeakers below the lower ring (0º). The 3D panning software also did not extend past the 0º elevation mark. By visual analysis of the data, it seems that bias from end-point effects influenced the angular error calculation on the 0º elevation results such that azimuth was the significant determining factor in our lower ring accuracy results. Future experiments should provide panning and loudspeaker arrays extending beyond the target destination in elevation.

6 CONCLUSIONS

6.1 General Conclusions and Further Discussion

The aim of this research was to determine the *influence of virtual environments on a listener's perceived sensation of motion* during the presentation of auditory motion and, *to determine if binaural crossfading is an efficient method for simulating horizontal auditory motion between two known sound source locations*. To facilitate this study towards a conclusion, simulations of auditory motion in virtual environments were designed to replicate measured, real-world auditory motion examples in real environments. Auditory motion was presented to listeners through binaural simulations and through a hemispherical loudspeaker array using established crossfading techniques. Visual simulations of the real environments were presented through conventional 2D video displays, and through a 360º video for VR HMD.

Based on a quantitative and qualitative analysis of the data provided through perceptual listening tests on the sensation of motion, localization tasks and participant comments, the following general conclusions can be made:

1. **The VAEs presented in this thesis did not significantly influence the listeners' sensation of motion**
2. **Binaural crossfading sufficiently translated the sensation of motion for broadband signals in the auditory motion simulations**

3. **Binaural crossfading provided a sensation of motion that is relatively unaffected by acoustic reflections**
4. **The presentation of virtual reality visuals via HMD did not impede the task performance of a motion-based 3D audio panning task as compared to the same task performed in a replicated real-world visual reference**
5. **Non-immersive video may reduce the sensation of motion**

These conclusions are made on the basis of experimental conditions within this study. The following sections will provide individual summaries of how these conclusions were derived and elaborate further on the analysis and design of the experimental methodologies.

6.1.1 Virtual Environments and the Sensation of Motion

The results indicate that the VAEs, presented under the conditions in this thesis, do not significantly influence the listeners' sensation of motion as compared to a real-world reference when evaluating horizontal auditory motion via binaural crossfading. In addition, their influence may slightly increase the sensation of motion as compared to the perception of real-world auditory motion (see Figure 34). For the first two experiments, simulations by condition were greater in all situations (audio and audio/video in anechoic and reflective environments) than the reference. In the VR experiment, the task performance related to one's ability to perceive a sensation of motion slightly increased over that of a real environment. It seems that the limited feature set of the virtual environments in this study allow the listener to focus more directly on the audio and visual stimuli under test. For instance, in the VR experiment, the virtual environment was seemingly empty other than for the modeled acoustic treatment on the external walls, lights overhead, floor below and

loudspeakers placed in mid-air at their respective positions. Conversely, in the real environment for the same experiment, the user was aware of the desk, computer equipment to perform the experiment and a higher resolution of reality than presented in VR—though the VR replica was still quite high in quality. Task duration (see Table 5 and Table 6) was also reduced for VR conditions in the VR experiment which suggests a quicker onset detection of motion in the virtual environment, leading to a greater sensation of motion. It's interesting to point out that for the speaker conditions in the VR experiment (Rs/Vs), there was no significant difference in localization accuracy or task duration. This concludes that replicated virtual environments must also include replicated sound source positions. When this occurs, the greatest one-to-one translation is possible for an audio panning task in virtual reality.

6.1.1.1 Influence on Localization

The positional results of the VR experiment reflect the reduced localization accuracy for sound objects appearing at height and above the head which is consistent with the literature [8] (see Table 5). However, clear end-point effects exist in the data at 90º N elevation (directly above the listener) because results show a near-perfect localization accuracy (see "position 9" in Table 5), which is in conflict to published results (see Figure 3). This issue most likely comes from a technical design fault in the VBAP panner, in which sound was "snapping" to the 90ºN position instead of moving freely past this point towards either side of the median plane.

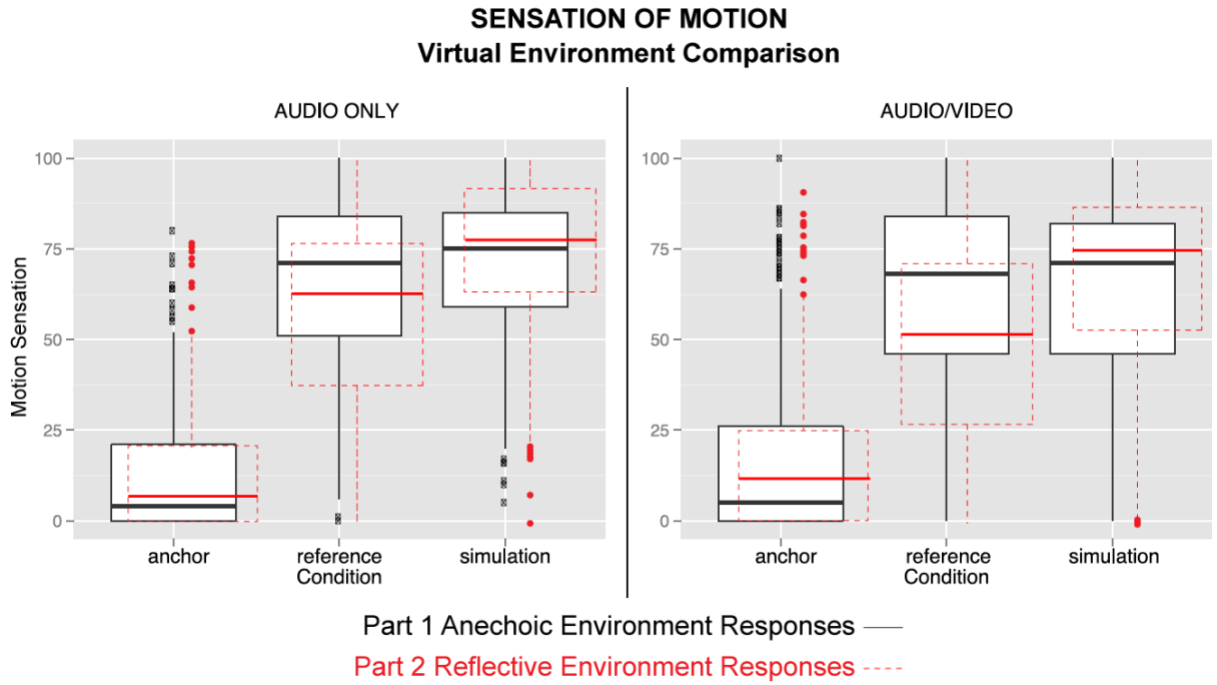

Figure 34: Boxplot showing the final comparison of grouped responses for motion sensation by condition from Part 1 and Part 2 of the binaural simulation experiments (black = Part 1 data, red = Part 2 data) grouped by audio only (left) and audio/video (right)

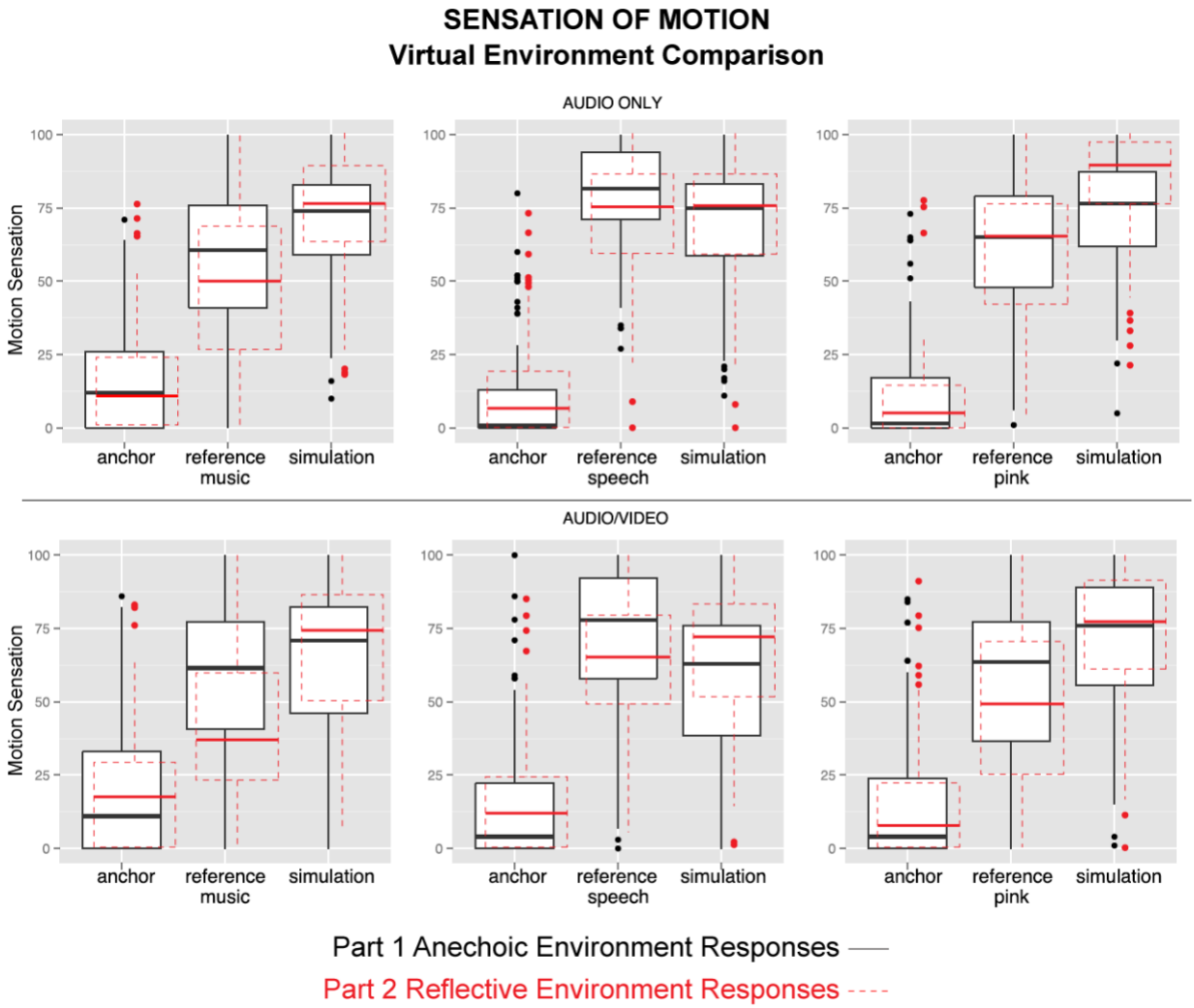

Figure 35: Boxplot of the final comparison of individual signal responses by condition from Part 1 and Part 2 of the binaural simulation experiments (black = Part 1 data, red = Part 2 data) grouped by audio only (top) and audio/video (bottom)

6.1.2 Sensation of Motion Presented through Binaural Crossfading in VAEs

For binaural motion simulations in VAEs, the technique of binaural crossfading for simulating

auditory motion provided measures of sensation of motion that were equal to, or greater than those reported through referenced recordings of binaural auditory motion. This can only be concluded for broadband signals as narrow band signals such as speech may require additional compensation to effectively translate the sensation of motion. As a stimulus in the first two experiments, the speech signal recorded greater measures of sensation of motion through the reference than through the simulation on all, except the audio/visual presentation in the reflective environment. It is thought that the strong reflective environment with the addition of inaccurate visuals, may have caused the sensation of motion to drop lower than the other reference results, in this instance. The greater performance of speech overall seems to suggest a natural tendency to recognize the binaural cues of a voice, since human sensitivity to voice is greatest among all signals [8].

The reflective environment heavily influenced the reference motion (see Figure 34 and Figure 35). However, it did not greatly influence the binaural crossfading simulation. This suggests that reference auditory motion recordings suffered from localization blur due to the prolonged, high energy early reflections present in the measurements of the reflective environment. Such strong reflections would have masked the onset binaural cues necessary for accurate localization at incremental positions along the motion path. Essentially, the reference recording seems to have captured a blurring effect as the dummy head was moved from one loudspeaker to the next. In stark contrast, the simulation included BRIRs with binaural cues from just three measured positions followed by the early and late reflections of the room at these positions. The three positions (and their cues) were then interpolated through the crossfading process, which essentially

turned down the volume of one set of cues and allowed the next to be heard in gradual succession. It is thought that the binaural cues were successfully maintained through this process, and since reflections did not continue to build up and hold high energy over time, a reduction of binaural masking occurred.

In general, video as a factor on binaural crossfading, did not significantly affect the sensation of motion, though in Figure 34, one can see a slight reduction in results. Perceptual cues in crossfading may be more effective at presenting a sensation of motion alone, rather than through the addition of cues from moving visual objects.

6.1.3 Influence of Virtual Environments on a Motion-Based Localization Task

The goal of the experiment in Chapter 5 was to apply the conclusions from the first two experiments into an application-based study in which a listener would require the sensation of motion to accurately perform a motion-based immersive audio production task. The experiment would serve to identify the influence of VR on auditory motion, with specific focus on the visual modality.

As discussed in Section 2.3, perception of auditory motion is relative to the sound source position and one's position in space (see Figure 9), which means that a simulated sound moving from one virtual position to another past a stationary listener would be perceived the same as simulated listener movement between two real sound sources in a free field (as in Chapters 3 & 4). The methodology developed for the VR experiment was conducted with this in mind. The

hemispherical loudspeaker array was chosen to maintain accurate and repeatable sound positioning and auditory motion over the course of the experiment. Conducting the same experiment using headphones would have required individualized HRTF databases to be captured and processed for each and every listening subject, head-tracking algorithms, additional localization pre-tests to determine the positional accuracy of such binaural simulations, and recalibration of visual elements to virtual sound sources per individual listening subject based on their individual head-maps (HRTF databases). Such a process would have been impractical based on the timeline and resources allocated for this study. Thus, the hemispherical loudspeaker array was chosen, which in turn helped to focus attention directly on the visual conditions. The same auditory motion technique from the first two experiments (crossfading technique), was used to control sound objects in the real auditory environment while presenting the listeners with a variable visual simulation through virtual reality and measuring this influence through their performance of accurate target localization.

The results indicate that in the design of VR audio productions, one should take care to replicate the cross-modal aspects of the virtual environment to the physical model such that virtual sound sources replicate the localized position of simulated or physical sound sources. When this occurred in this study, the presentation of virtual environments through VR did not negatively influence one's performance of a motion-based audio task requiring the sensation of motion. In fact, results suggest it made the task easier in terms of speed and accuracy, as Section 5.6 points out.

6.1.4 Influence of Visuals on Results

Based on the audio/video results in Chapters 3 and 4, it is important to give the viewer of immersive audiovisual productions a full field of view when presenting them with immersive/3D audio in virtual environments. The VR experiment further proved this point; by allowing the listening subjects to use head movements to localize the probe sound to the target sound, they were much more accurate in localization than in the restrained head simulation used in the first two experiments. Comments from participants in the first two experiments repeatedly stated that the visual environment did not seem to match what they were hearing. It seems that in the reflective environment (see Figure 34), the strong localization blur caused by the acoustic reflections in the measured reference had combined with negative aspects of the limited field of view video to cause the large difference in sensation of motion results. This impact could have also been due to the method for capturing video motion (see Section 3.5.2) as a reference for simulating replicated auditory motion paths. However, the results of the simulation shown in Figure 34 and Figure 35 suggest that the video capture method did well to pair with the auditory motion in the simulation. In the VR experiment, localization accuracy was decreased when visual reference to loudspeakers was limited or when visuals were taken away altogether, confirming expected results in this area [88] [106] [107]. Overall, the results of visual influence suggest that head-tracking should be included in binaural motion simulations in virtual environments, and full field of view should be provided to the viewer of VR productions.

6.2 Recommendations for Future Work

Sensation of motion is just the first step in understanding how real-world motion-based audio productions translate to virtual reality. While this study provided a thorough analysis of the sensation of motion, several perceptual factors and challenges still need to be addressed in future research. This section will provide an overview of these points and propose recommendations for how they could be addressed.

6.2.1 Additional Factors

One of the main goals of an accurate simulation is to replicate all of the perceptual factors within a referenced real-world event. The methodology for the first two experiments within this thesis examined the sensation of motion on a metric of perceptual detectability. Further work would be required to determine the weight of this metric against the total perception of the auditory motion under test. In addition to sensation of motion, investigations could be made into the multiple factors that occur during the reference capture of auditory motion. These factors, along with suggested literature include: duration and velocity [28] [63], *minimum audible movement angle* (MAMA) [80], perceived path (direction) of motion [156], head movements [111] [157], source directivity [71], and detection of self-motion vs. apparent motion [28] [59]. Analysis of these factors would support a broader conclusion on the necessary components required for accurate binaural simulation of auditory motion.

6.2.2 Perceptual Evaluation Methodology

As this thesis has illustrated, in order to understand how auditory motion perception translates

from real to virtual environments, a standardized method is needed for the comparative evaluation of the sensation of motion (and its associated percepts) in real and virtual environments. Methods exist for the evaluation of audio quality [113] [77] and for measuring perceptual thresholds of auditory motion percepts [28]. Therefore, these methods, along with those designed for this thesis should serve as guidelines and influences for future studies, validation of results, and work towards an established method for subjective assessment in this area of research.

6.2.2.1 Limitations of the "Sensation of Motion" Metric

The "sensation of motion" metric was defined for this study to encompass multiple factors of motion phenomena to serve as a precursory report of the quality of simulated auditory motion in virtual environments and virtual reality. Little work had been done to support quality of motion perception with a focus on the exact replication of auditory motion in virtual reality. Most studies simplified the virtual environment to focus on more precise auditory measures. However, in this study, it was hoped that a simplified measure of motion sensation could lead to further, more advanced studies by first focusing on a replication methodology that would support an analysis of equivalence and quality between a reference and a simulation. However, following the completion of the studies, it became apparent that equivalence results could have been further supported through a perceptual difference test. As a recommendation for future work, the author has elaborated on the areas of study that should be refined for more precise analysis and conclusions.

6.2.2.1.1 Classifications of Motion Sensation

As demonstrated in Mathiesen's landmark study [55], one should be aware that motion sensation

can be described by multiple phenomena, for example: apparent motion vs. self-motion, as well as the descriptive variables of each. Mathiesen's methods sorted listener responses by distinct classes of motion perception to properly define the influence of such variables on the overall perception of motion. By simplifying the "sensation of motion" variable (in this thesis) into a single comparative rating evaluated by a 5-point scale on the basis of quality ("Bad" to "Excellent"), the author overlooked the potential influence of additional factors of the motion perception phenomena on the experimental results. While continuous quality rating scales are sufficient for comparisons of static (motionless) audio quality, this process unintentionally limited the depth of analysis for the sensation of motion to an all-encompassing value of equivalence, thus providing an incomplete understanding of the underlying motion sensation percepts. Limitations caused by the singular "sensation of motion" metric can be found in the optional written comments by assessors (see Section 3.8.2.2). An example of such confusion and its potential influence on the results becomes apparent in the comment from Assessor V016 where mention of a greater "physical sense of movement" was experienced in the voice stimuli. The initial analysis in Section 3.8.2.1 references this comment in support of the voice stimuli exhibiting greater sensation of motion ratings due to potential sound source familiarity. However, a greater depth of knowledge could have been provided had the experiments included rating scales for each descriptive class of motion perception that influenced the sensation of motion responses. In this way, the physical motion experienced by Assessor V016 could have been further evaluated, potentially offering further insight (along with other classifiers) into why the reference voice exhibited higher sensation of motion ratings.

For example, in Chapter 3 and Chapter 4, the sensation of motion responses cannot be evaluated or described by the *transition* of motion experienced by the assessors during the motion events. Perceptual tests did not require the listeners to report the level at which motion was perceived to be smooth, jumpy, or inconsistent, but comments do indicate that some of these descriptors influenced the sensation of motion responses. While comments cannot provide exact measures of influence, they do shed light into the motion phenomena that was experienced. Upon full review, it was found that the motion experienced in both experiments was commonly reported as having smooth transitions from one source position to another. Comments also suggest the motion was not exactly following a straight path and could present inconsistent motion to the visual reference, but there was no mention of sound jumping from one position to the other (similar to what Mathiesen described as "two-ness" or the perception of two distinct auditory sources) [55]. These comments suggest that the binaural simulation provided a smooth transition of motion, but the aforementioned factors may have influenced the accuracy of the simulated motion path—an important consideration for anyone attempting to simulate reality.

Without metrics to describe self-motion vs. apparent motion, as well as the perceived transition of motion, an analysis of the "sensation of motion" can only be reported through conditional limitations and thus, it is imperative that future studies include such classifications in the design of motion sensation experiments.

6.2.2.1.2 Potential Influence of Signal Coloration and Other Bias in Perceptual Analysis

For Chapters 3 and 4, it should be noted that potential signal coloration may have existed in the

comparison of simulation stimuli to reference stimuli. This might exist if the simulation techniques failed to accurately replicate the acoustic conditions of the reference recordings. An inability to accurately replicate the reference signal would have likely presented signal coloration in the simulation in the form of changes in phase and the resulting frequency response, changes in intensity and sound field imaging, and differences in additional acoustic phenomena—especially those related to acoustic reflections. Even though the stimuli were auditioned through a double-blind presentation, unintentional signal coloration may have caused significant differential weighting or preference in one stimulus to another and could have biased the evaluation of the sensation of motion towards the signal coloration attributes. This weighting may have occurred when an assessor evaluated a first stimulus that contained significant sonic attributes due to signal coloration, and then experienced changes in coloration in a subsequent stimulus. Such biases are known to occur through the effect of sequential contraction, when strong changes in sonic attributes are present [113] [158]. It is also possible that additional looming bias may have been present due to the simplified method of simulating auditory motion [159]. Looming bias defines the phenomenon of sound sources becoming more salient as sound intensity increases, than for sound intensities decreasing. The crossfaded signals in the simulation could have presented greater ramps in intensity as compared to the reference, or vice versa, thus causing one of the conditions to be more salient and more preferred for exhibiting greater sensation of motion. Though this is an unsupported and indirect assumption, it could be possible that looming bias extended itself into the comparison of these intensity ramps, especially in the case of potential timing differences between conditions.

Furthermore, the design of the assessments and the limitation of a single response metric could have presented the potential operation of a demand characteristic on the participant's responses in which the hypothesis was surmised by participants through the stimulus assessment method [160]. This study took steps to try to eliminate such biases by adopting the MUSHRA technique but given the limited number of stimuli to evaluate (3), the MUSHRA technique may have failed to prevent clear and unbiased stimulus discrimination. The experiments in Chapters 3 and 4 failed to conduct a comparative analysis by the participants to determine the exact influence due to signal coloration, if any, between stimuli conditions. A more thorough method of evaluation for the sensation of motion might include perceptual analysis of the similarities between key attributes presented in the different conditions such that multi-variable analysis could inform the overall metric of sensation of motion without directly posing the question to participants. Instead, the researcher could derive correlations between specific attributes which describe the sensation of motion response. This process would determine what perceptual differences or weights may be present among these specific attributes and may yield a greater understanding of how these differences might alter the sensation of motion from reference to simulation. The operation of a demand characteristic would be significantly prevented through this process.

Despite the potential for such biases to occur, it should be noted however, that an extensive process was conducted through the capture method to replicate the full acoustic response in all simulations, as mentioned in Section 3.5. This detailed method should have presented limited signal coloration at all static positions. It is thought that any coloration presented in this experiment would have

been weighted more toward the method of motion simulation through signal interpolation (crossfading technique) than through the technique for capturing the BRIRs

6.2.3 Considerations for Auditory Motion Interpolation

The technique to simulate auditory motion in the first two experiments was performed through linear crossfading. The results suggest that the linear crossfading technique may limit the binaural cues that would otherwise be present in a full simulation of motion. Therefore, studies conducted in reflective environments might benefit from the use of more realistic interpolation techniques such as logarithmic crossfading, VBAP and other advanced resolution techniques reported in [71]. It is recommended that future studies provide greater resolution between interpolated positions to match the reported findings on MAA and MAMA within the literature (see Section 2.1.2.6) so that motion interpolation can be confirmed to perceptual thresholds of detection. These interpolation methods could also be evaluated further for their accuracy through a corresponding evaluation of MAMA. In this case, MAMA results might indicate or confirm deficiencies in the angular resolution of the sensation of motion for a given velocity.

6.2.4 Virtual Reality and Head-Tracking

Due to time and technical constraints, the design of the first two experiments (see Section 3.4.3) was not replicated in VR as initially hoped. In addition, the last experiment was not performed through headphones using an HRTF database. These areas could be explored further. Simulation of the hemispherical array and the translational motion apparatus within VR, along with auditory motion presented through headphones may provide interesting results on the influence of reflective

environments. Head-tracking was not used in this thesis and could be applied to each experiment to compare the free-head movement results from this thesis to fixed-head and head-tracked results. To support future studies in this area, the anechoic VR environment and the modeled loudspeakers have been donated to the Sound Recording Program at McGill University in the form of a Unity software VR project.

7 APPENDIX A

7.1 Simulation Detail for Chapter 3

Important information relating to the design and implementation of the experimental methodology of Chapter 3 could not be provided in the originally published versions of the article due to length restrictions. Since this article has been included in the greater body of work that makes up this thesis, further clarification on the methodology is required. This section provides addition figures and tables as referenced in Chapter 3. Further clarification is provided through graphs of the 6 BRIR measurements from the semi-anechoic environment. Additional tables and figures detail the timing of the crossfading process.

7.1.1 BRIR Measurements from the Semi-Anechoic Environment

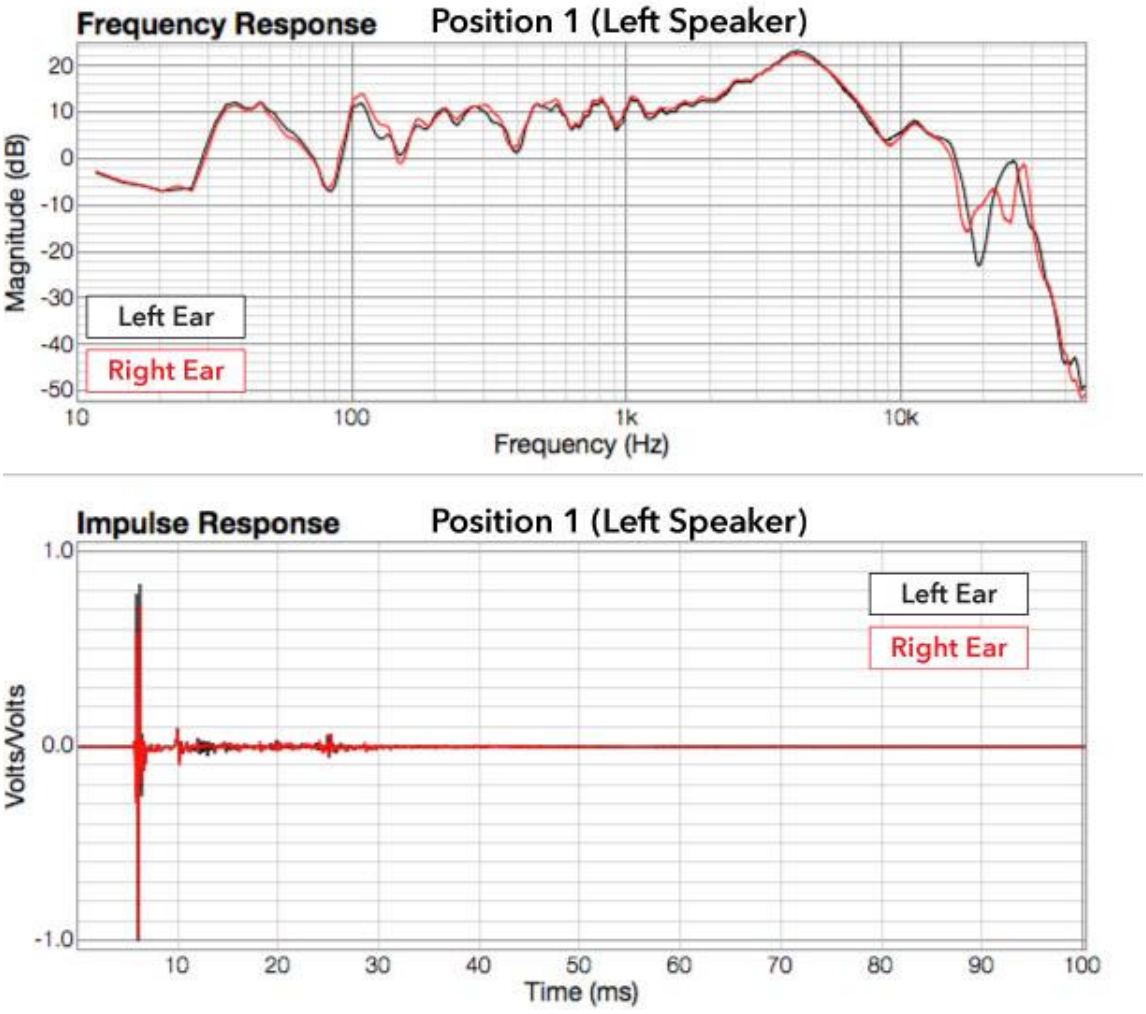

Figure 36: BRIR/frequency response for the left loudspeaker measured at position 1 (left) in the semi-anechoic environment

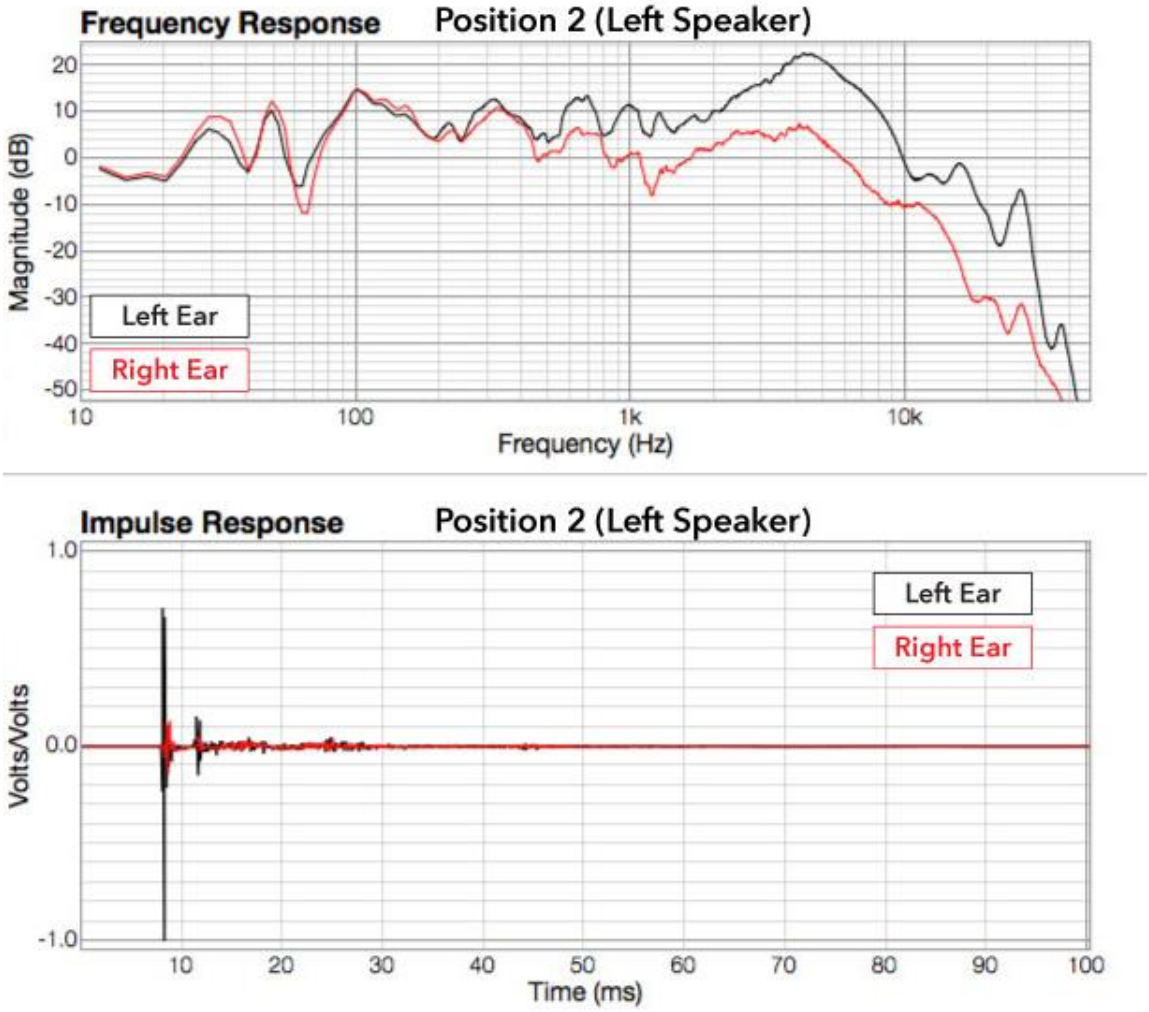

Figure 37: BRIR/frequency response for the left loudspeaker measured at position 2 (right) in the semi-anechoic environment

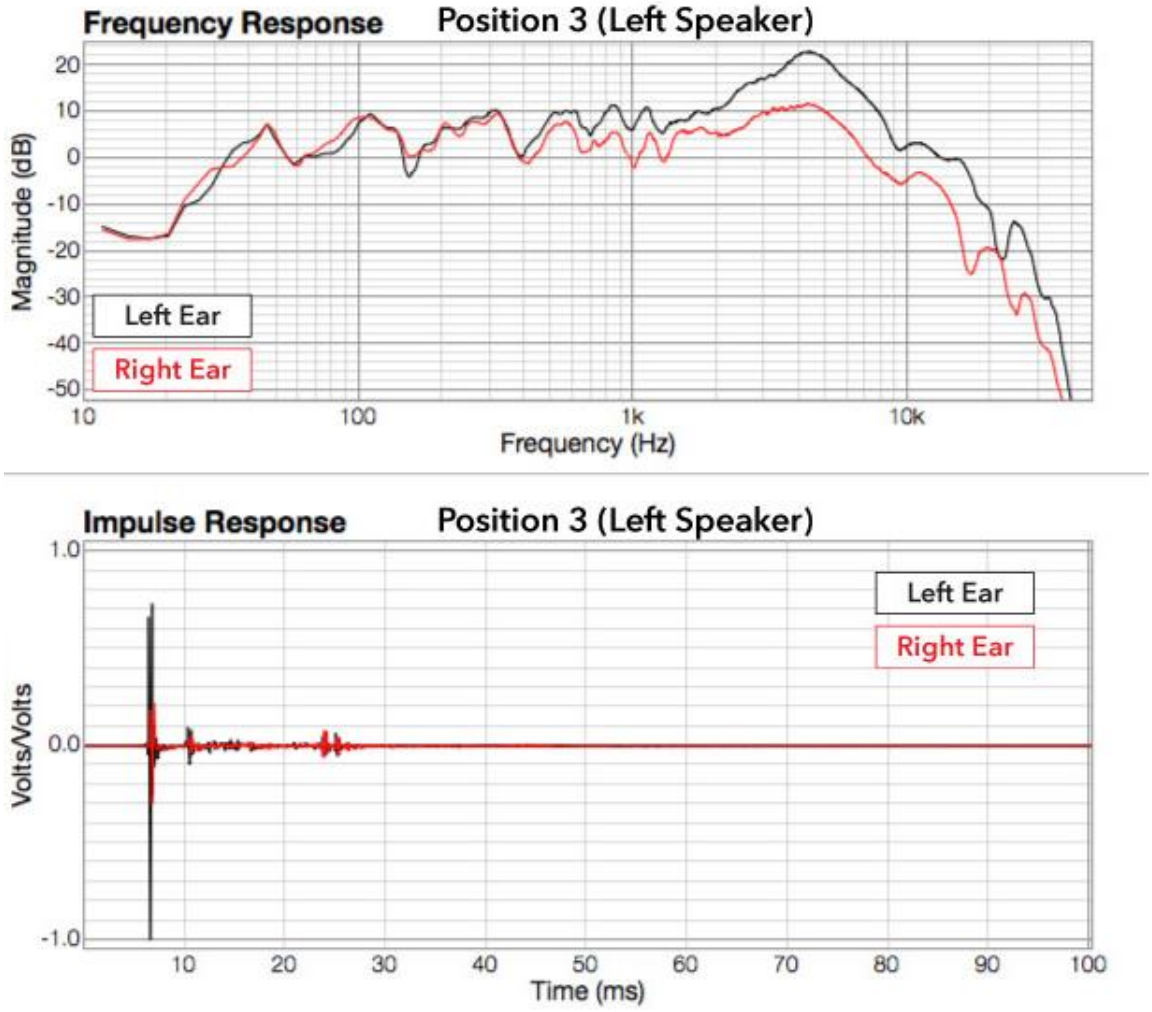

Figure 38: BRIR/frequency response for the left loudspeaker measured at position 3 (center) in the semi-anechoic environment

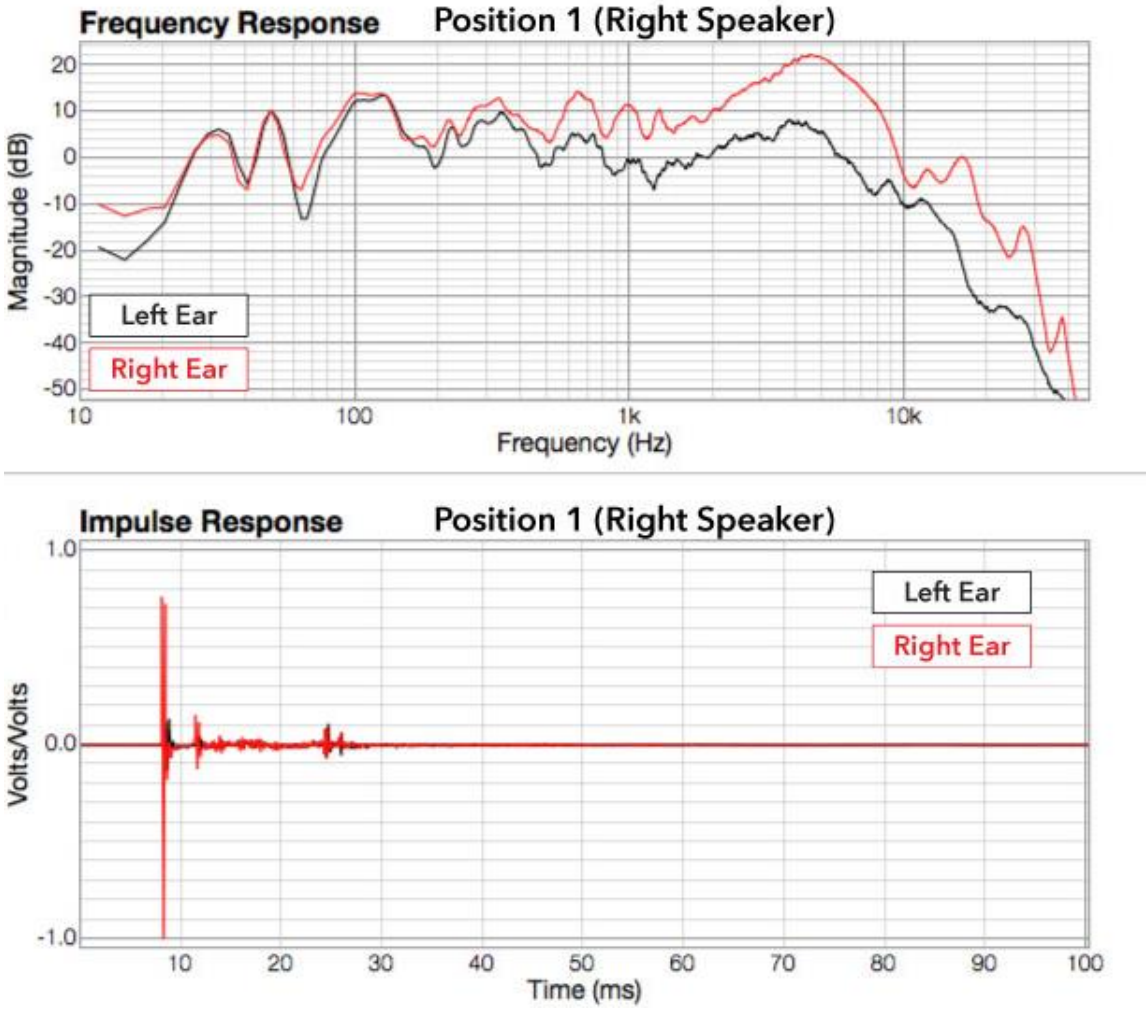

Figure 39: BRIR/frequency response for the right loudspeaker measured at position 1 (left) in the semi-anechoic environment

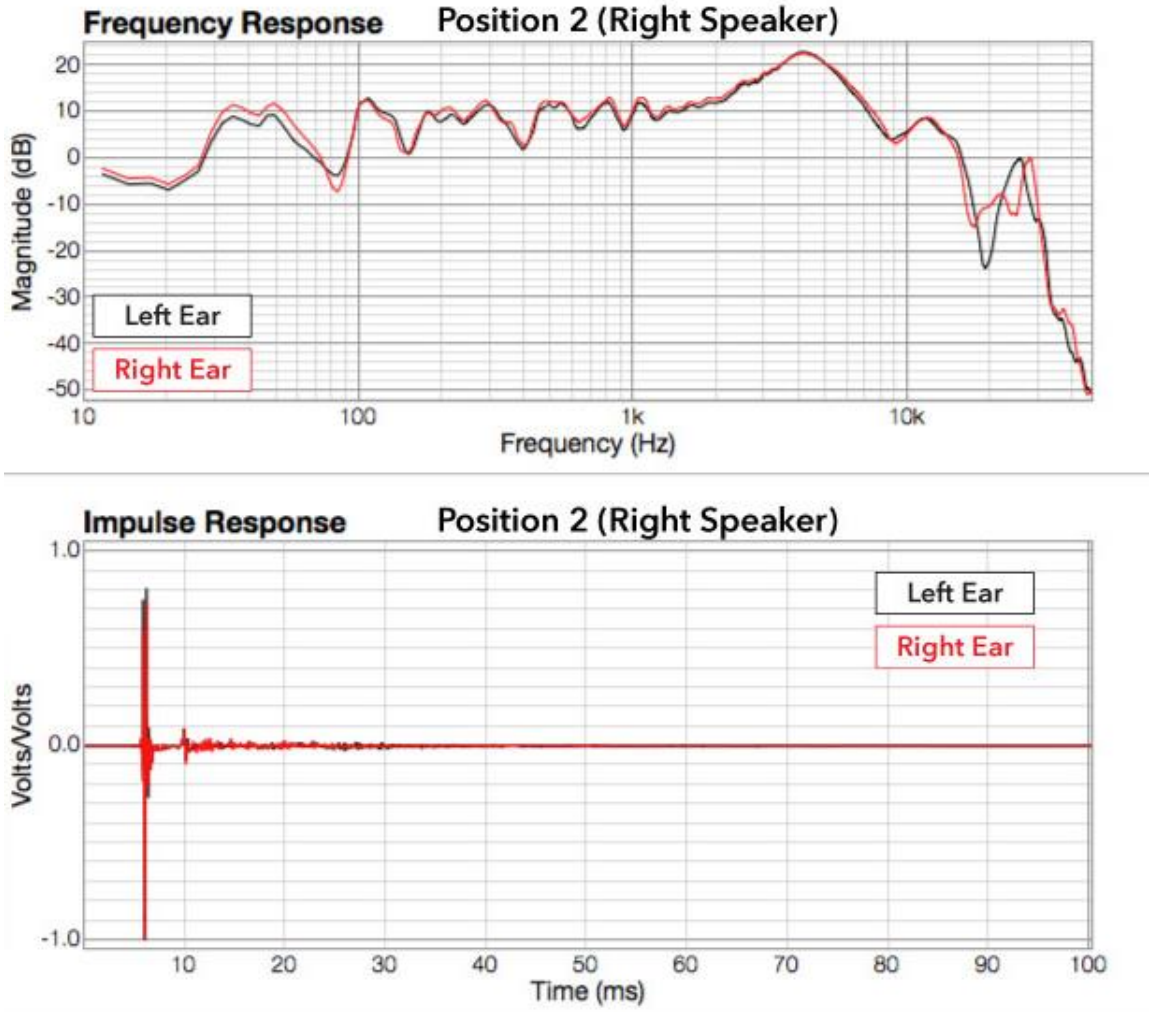

Figure 40: BRIR/frequency response for the right loudspeaker measured at position 2 (right) in the semi-anechoic environment

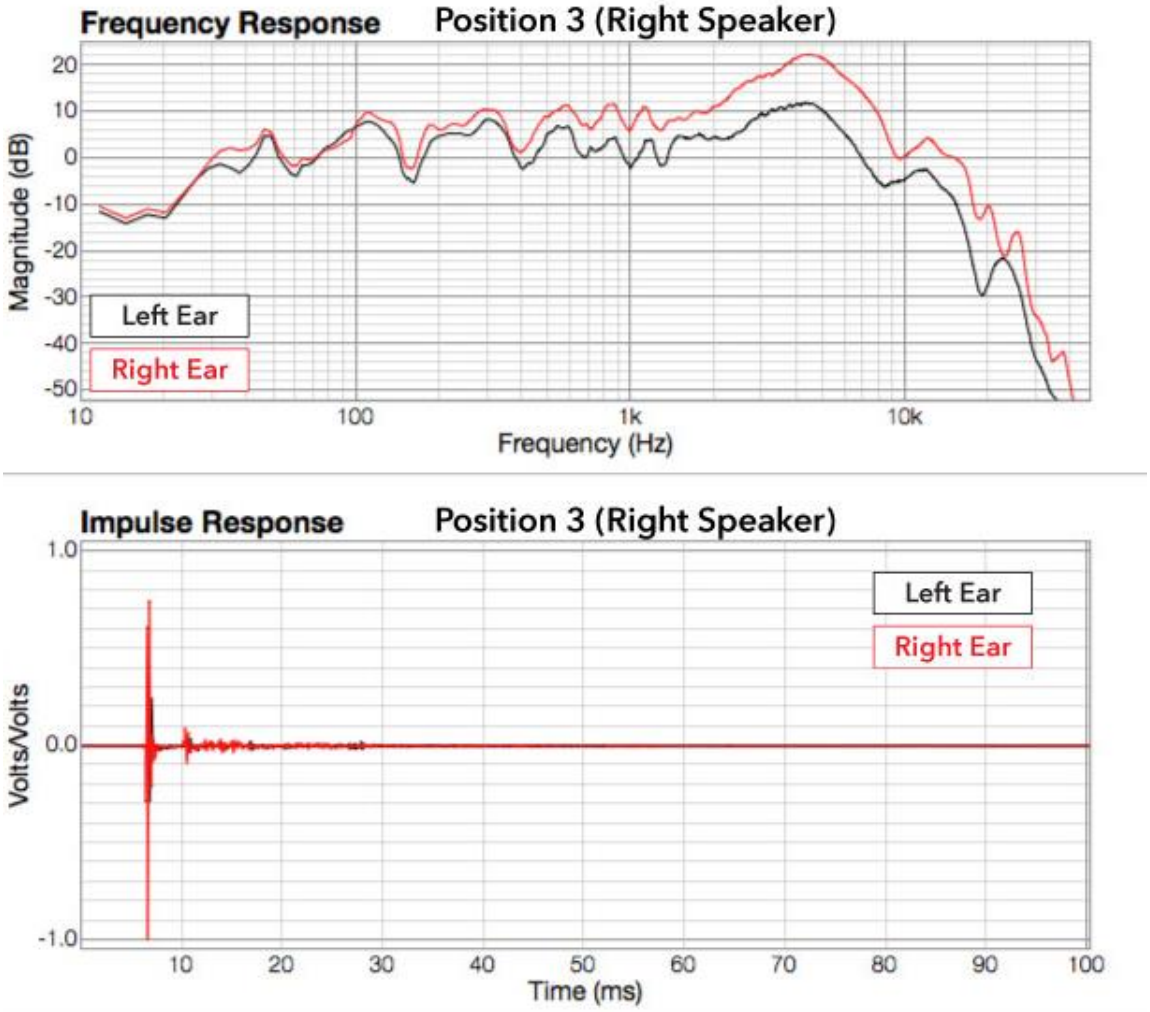

Figure 41: BRIR/frequency response for the right loudspeaker measured at position 3 (center) in the semi-anechoic environment

7.1.2 Crossfade Timing Detail (Chapter 3)

In Figure 42 below, the process of crossfading is detailed through illustration of the linear crossfading ramps for each of the output volumes for P1, P2 and P3 binaural audio simulation

stimuli. This process was segmented to align with 9 reference time intervals in each full motion pass. Figure 42 shows the crossfade design of the left to right motion pass starting at P1 and ending at P3. The opposite ramping technique was used to simulate motion in the right to left direction. Ramp time intervals (t_n) indicate the timing applied to the ramp between markers to match the video reference of binaural motion recording. Starting at t_1=0, as P1 audio output decreased, P3 audio output simultaneously increased at an equal rate based on the ramp time intervals. The process was repeated from P3 to P2 until stopping at full volume for P2. This crossfading process was repeated for the simulations performed in Chapter 4. Please refer to Appendix B for further detail of that process.

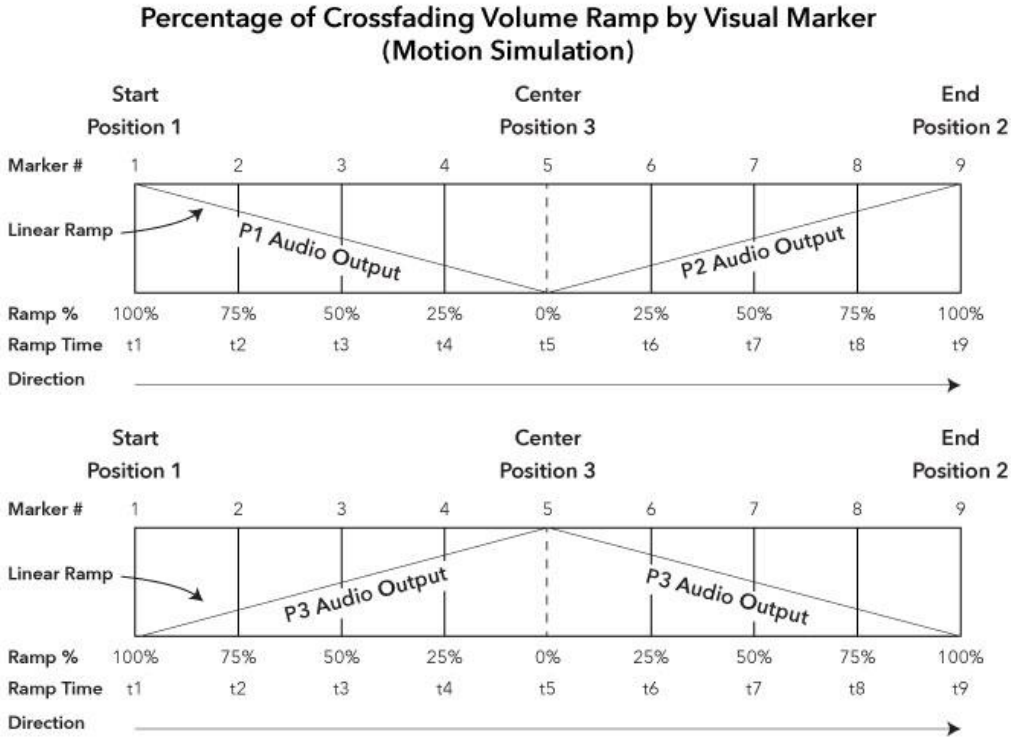

Figure 42: Detail of the linear ramps applied to the output volumes of P1 (left), P3 (center), and P2 (right) binaural audio simulation stimuli to simulate motion in the right direction.

A detailed crossfade ramp timetable for the experiment performed in the semi-anechoic experiment of Chapter 3 is presented in Table 7 where ramp time intervals are listed for each marker along the motion path. For the sake of simplicity, marker 1 should always be considered the start of the motion and marker 9 as the end (for either direction). Position 1 has a time of 0 ms to represent the start of motion. Total time is given for comparison of elapsed time in each motion path.

Table 7: Crossfade ramp timetable for the binaural auditory motion experiment performed in the semi-anechoic environment. All values are given in milliseconds.

Stimulus & Direction	Chapter 3 - Ramp Time Intervals by Marker (time in ms)									
	1	2	3	4	5	6	7	8	9	**Total**
Speech L2R	0	1216	946	853	962	1005	976	1058	1696	**8712**
Speech R2R	0	1024	935	874	938	1056	949	1152	1494	**8422**
Music L2R	0	1053	901	845	968	960	1021	1040	1621	**8409**
Music R2L	0	1113	877	836	954	959	1002	1056	1522	**8319**
Pink L2R	0	1181	1042	901	997	933	1013	984	1418	**8469**
Pink R2L	0	1181	938	925	941	901	936	1061	1392	**8275**

8 APPENDIX B

8.1 Simulation Detail for Chapter 4

Important information relating to the design and implementation of the experimental methodology of Chapter 4 is listed. This section provides addition figures and tables as referenced in Chapter 4. Further clarification is provided through graphs of the 6 BRIR measurements from the reflective environment. Additional tables and figures detail the timing of the crossfading process.

8.1.1 BRIR Measurements of the Reflective Environment

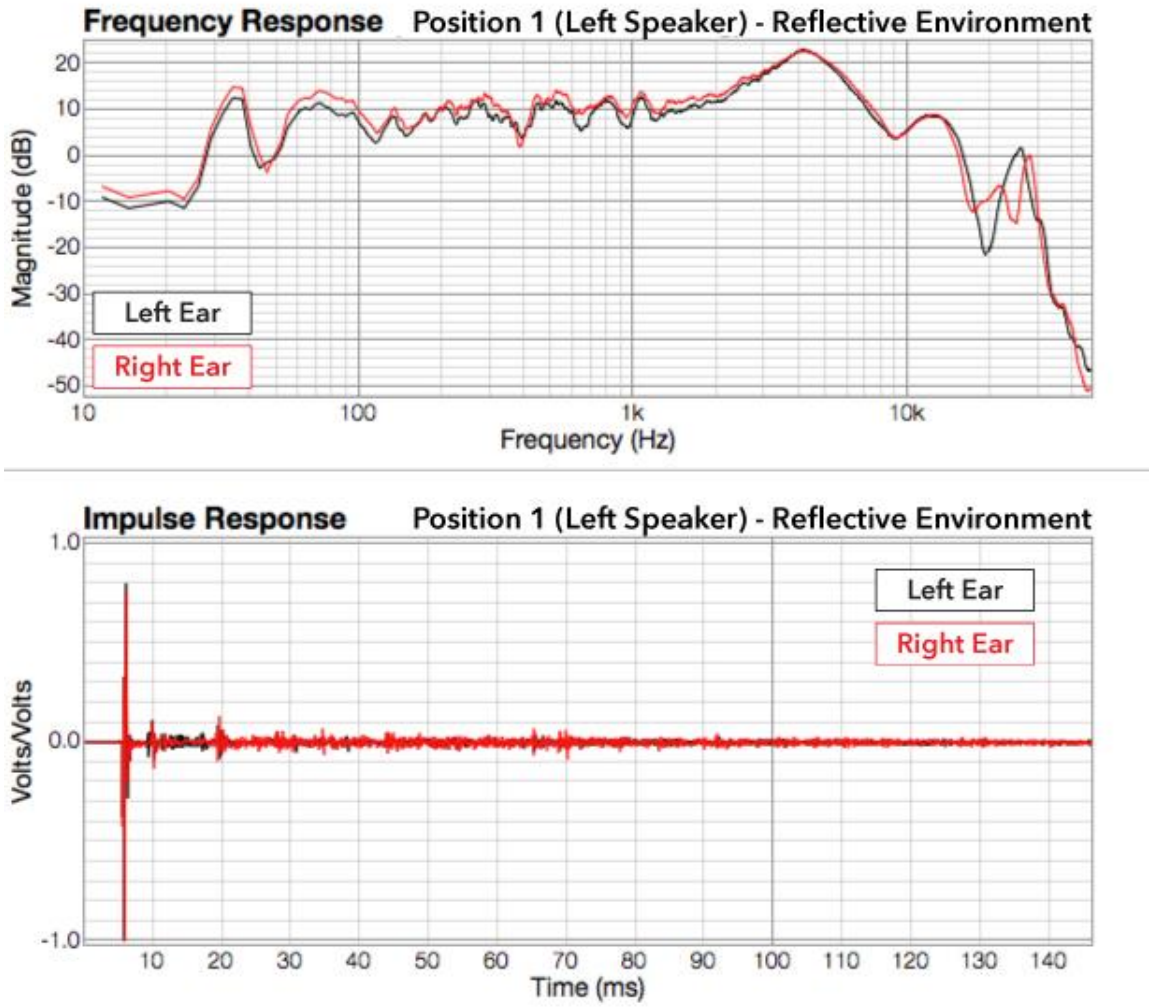

Figure 43: BRIR/frequency response for the left loudspeaker measured at position 1 (left) in the reflective environment

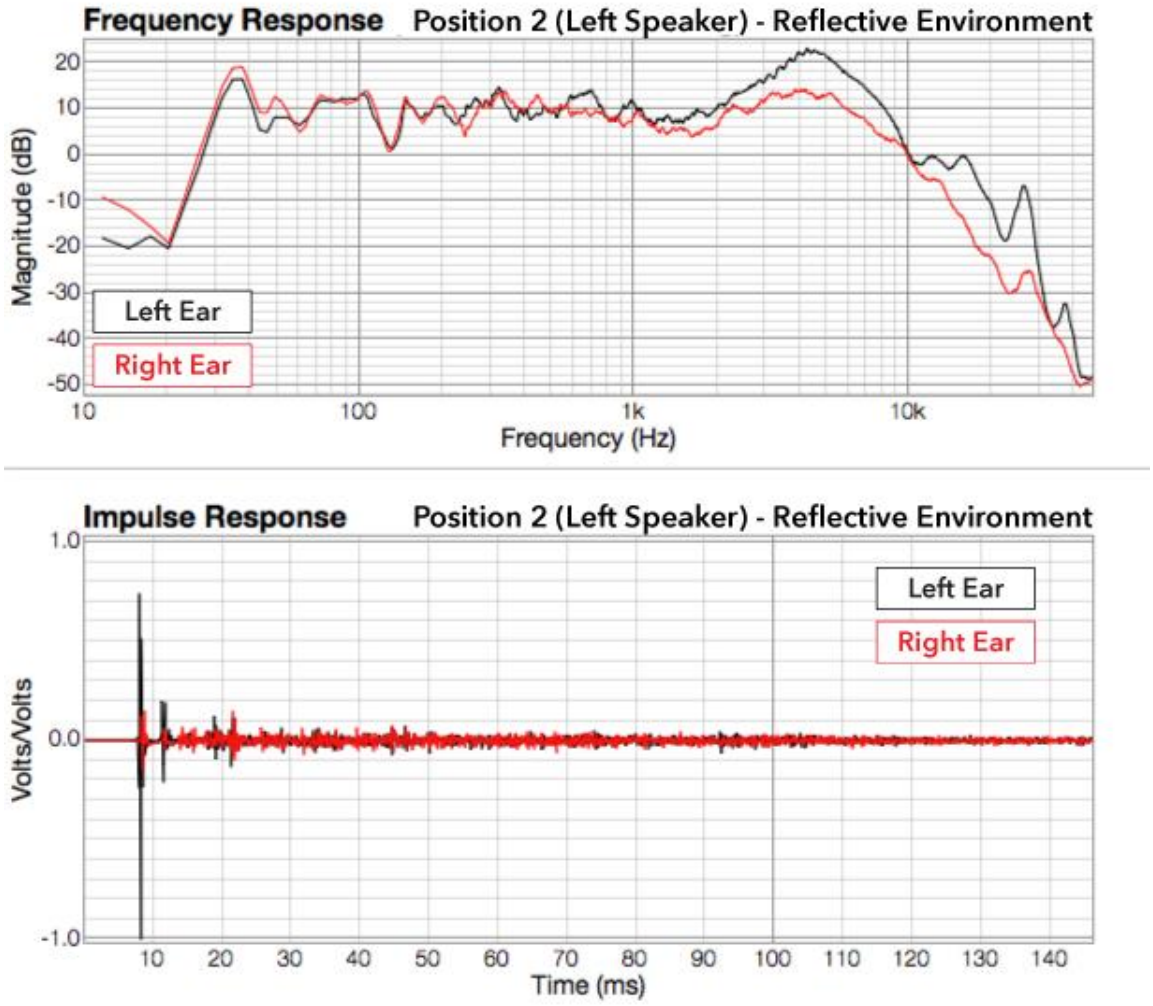

Figure 44: BRIR/frequency response for the left loudspeaker measured at position 2 (right) in the reflective environment

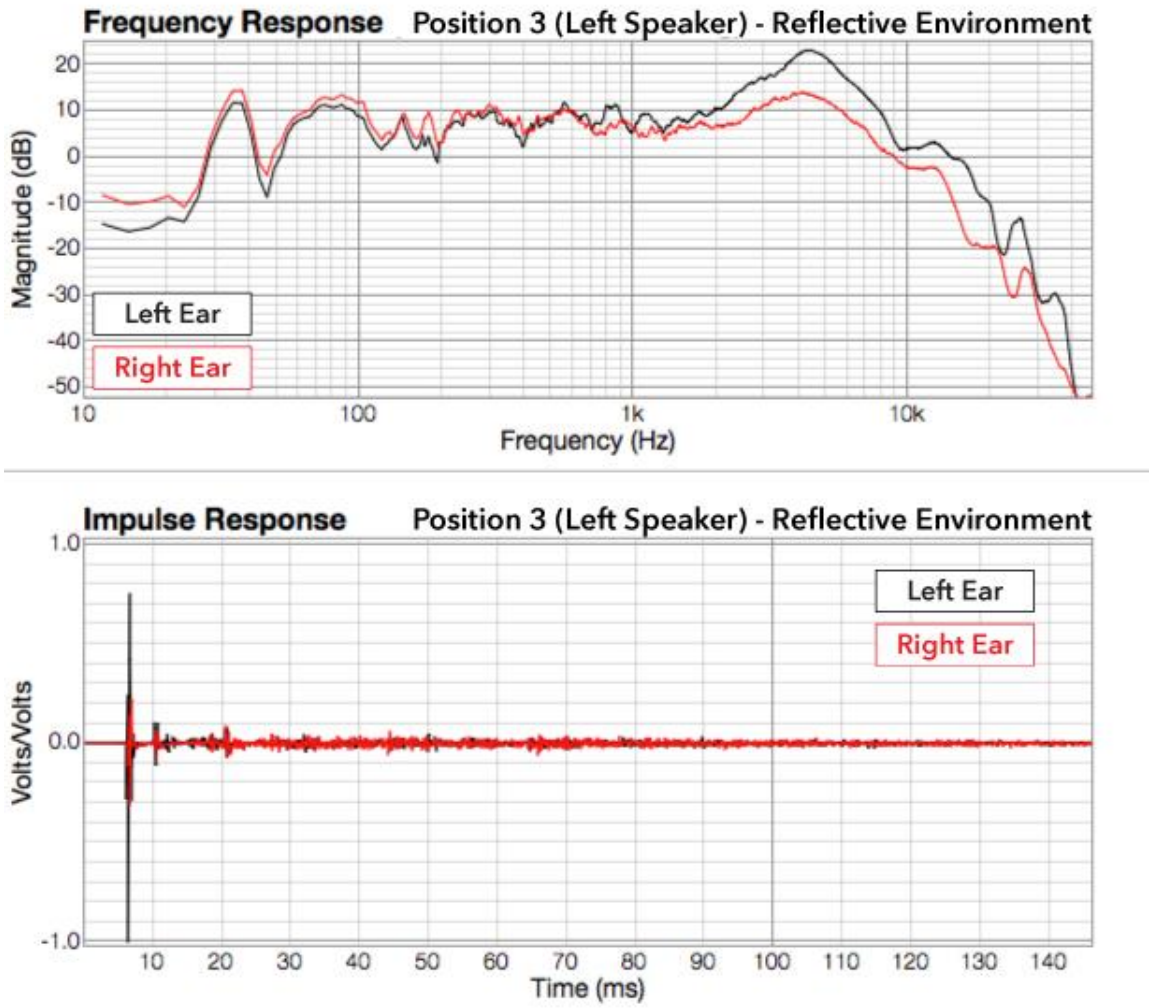

Figure 45: BRIR/frequency response for the left loudspeaker measured at position 3 (center) in the reflective environment

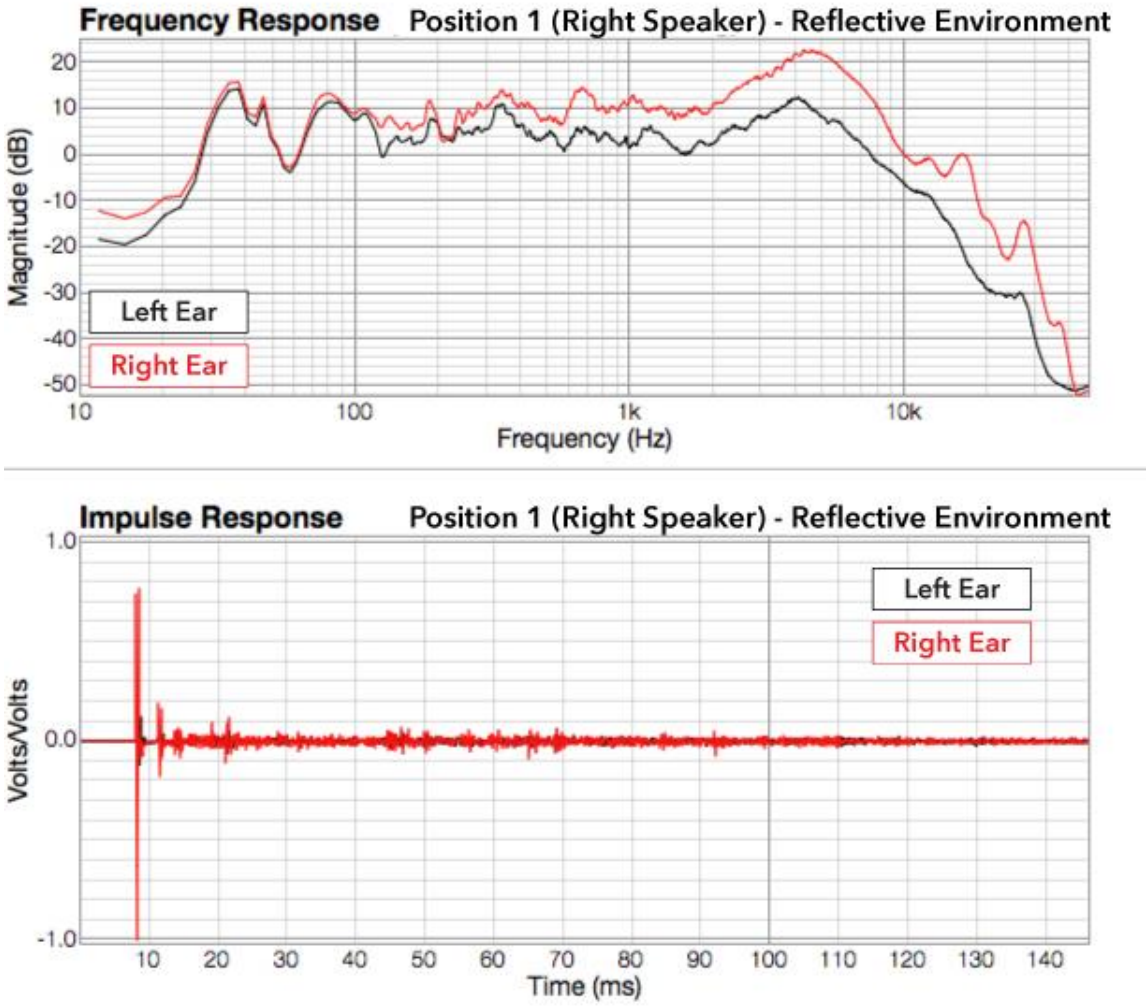

Figure 46: BRIR/frequency response for the right loudspeaker measured at position 1 (left) in the reflective environment

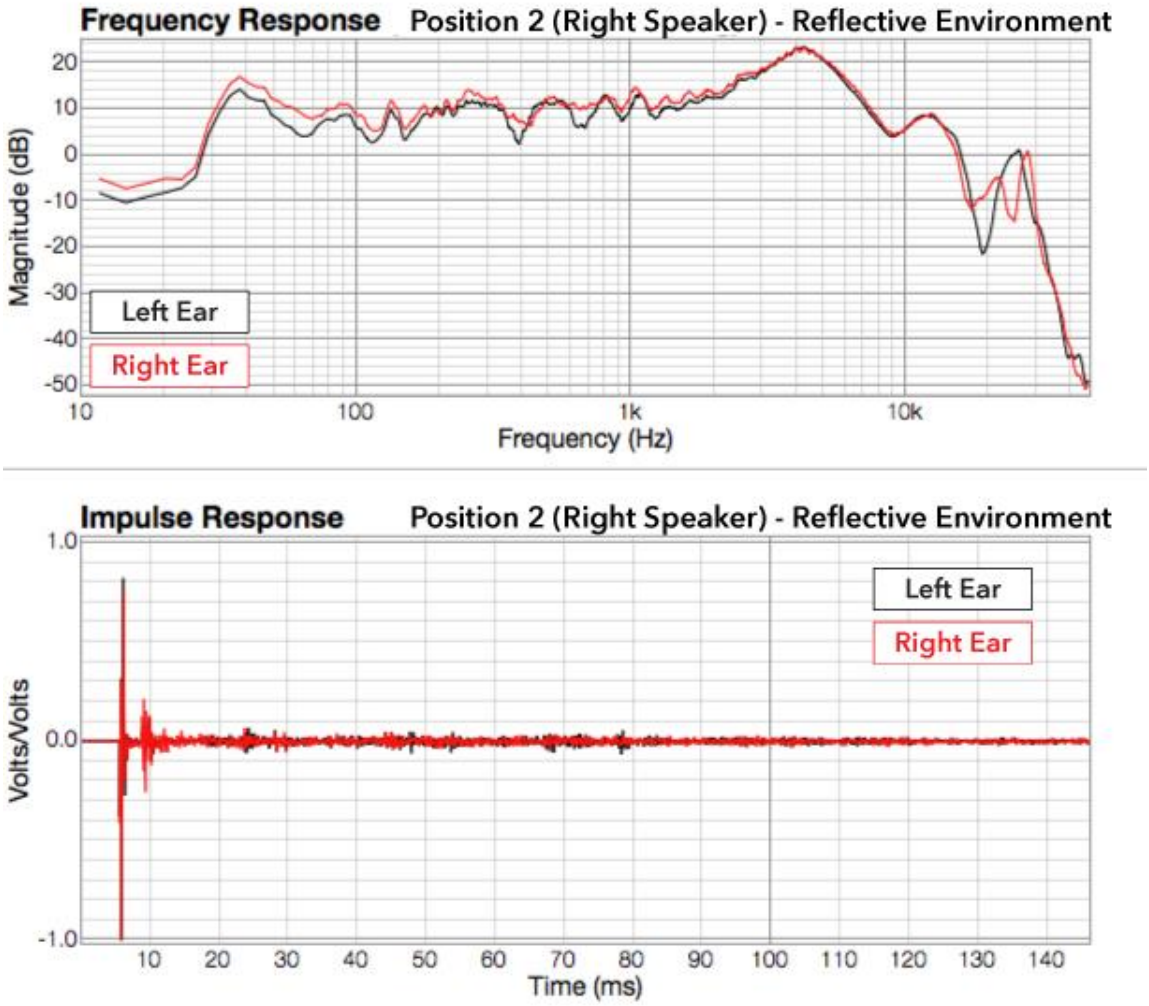

Figure 47: BRIR/frequency response for the right loudspeaker measured at position 2 (right) in the reflective environment

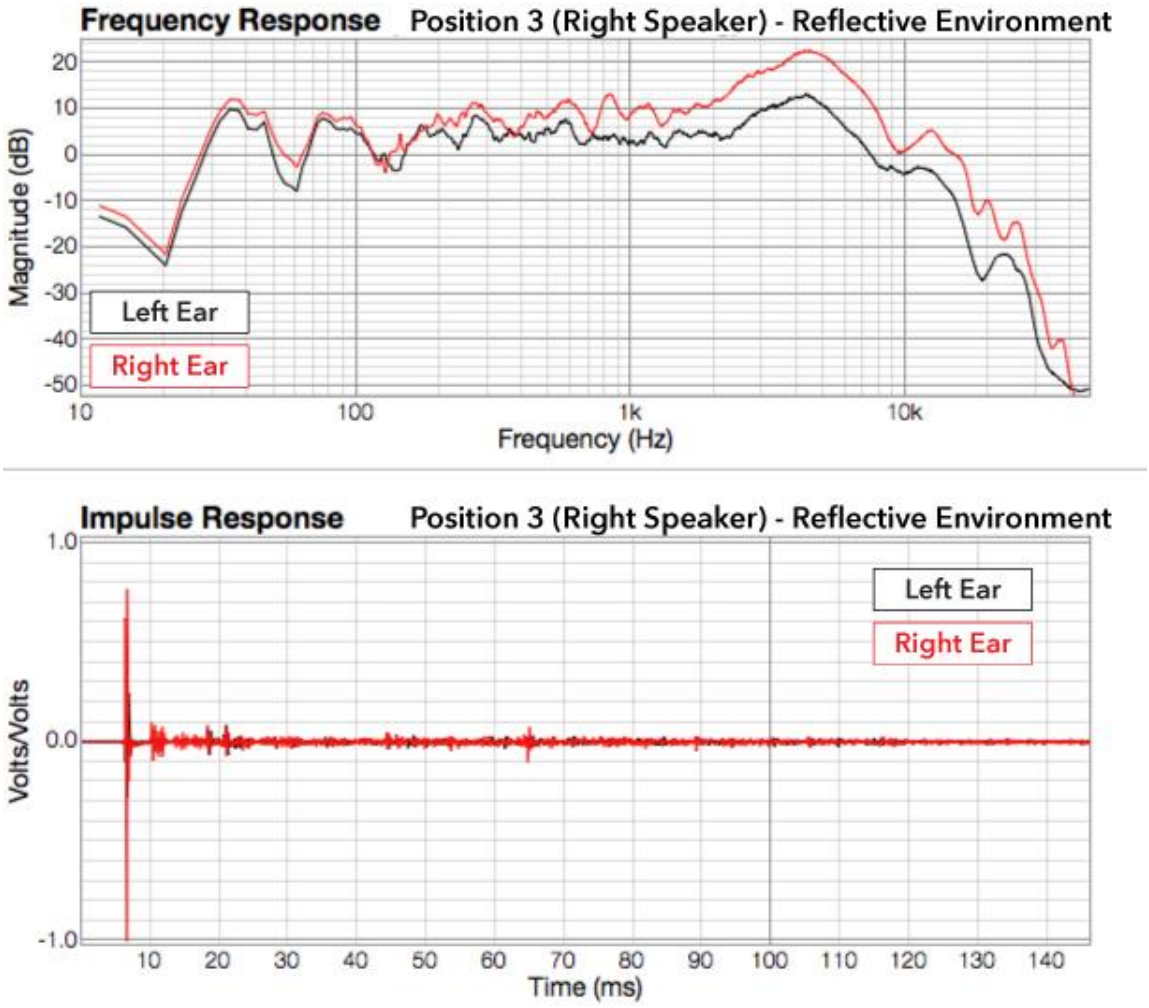

Figure 48: BRIR/frequency response for the right loudspeaker measured at position 3 (center) in the reflective environment

8.1.2 Crossfade Timing Detail

A detailed crossfade ramp timetable for the experiment performed in the reflective (reverberant) environment experiment of Chapter 4 is presented in Table 8.

Table 8: Crossfade ramp timetable for the binaural auditory motion experiment performed in the reflective environment. All values are given in milliseconds.

Stimulus & Direction	Chapter 4 - Ramp Time Intervals by Marker (time in ms)									
	1	2	3	4	5	6	7	8	9	**Total**
Speech L2R	0	1134	934	900	934	934	967	1167	1102	**8072**
Speech R2R	0	1334	933	968	967	901	967	1101	1234	**8405**
Music L2R	0	1001	967	901	867	901	900	1101	1434	**8072**
Music R2L	0	934	1067	1034	1001	934	900	1101	1268	**8239**
Pink L2R	0	1097	1068	1001	1000	967	968	1168	1200	**8469**
Pink R2L	0	1020	1001	867	968	1000	900	1001	1501	**8258**

8.2 Additional Setup Detail for the Reverberant Environment

Due to page restrictions during the original publication of the manuscript that makes up the bulk of Chapter 4, visual reference on the experimental setup in the reverberant environment could not be provided. Figure 49 now provides this information.

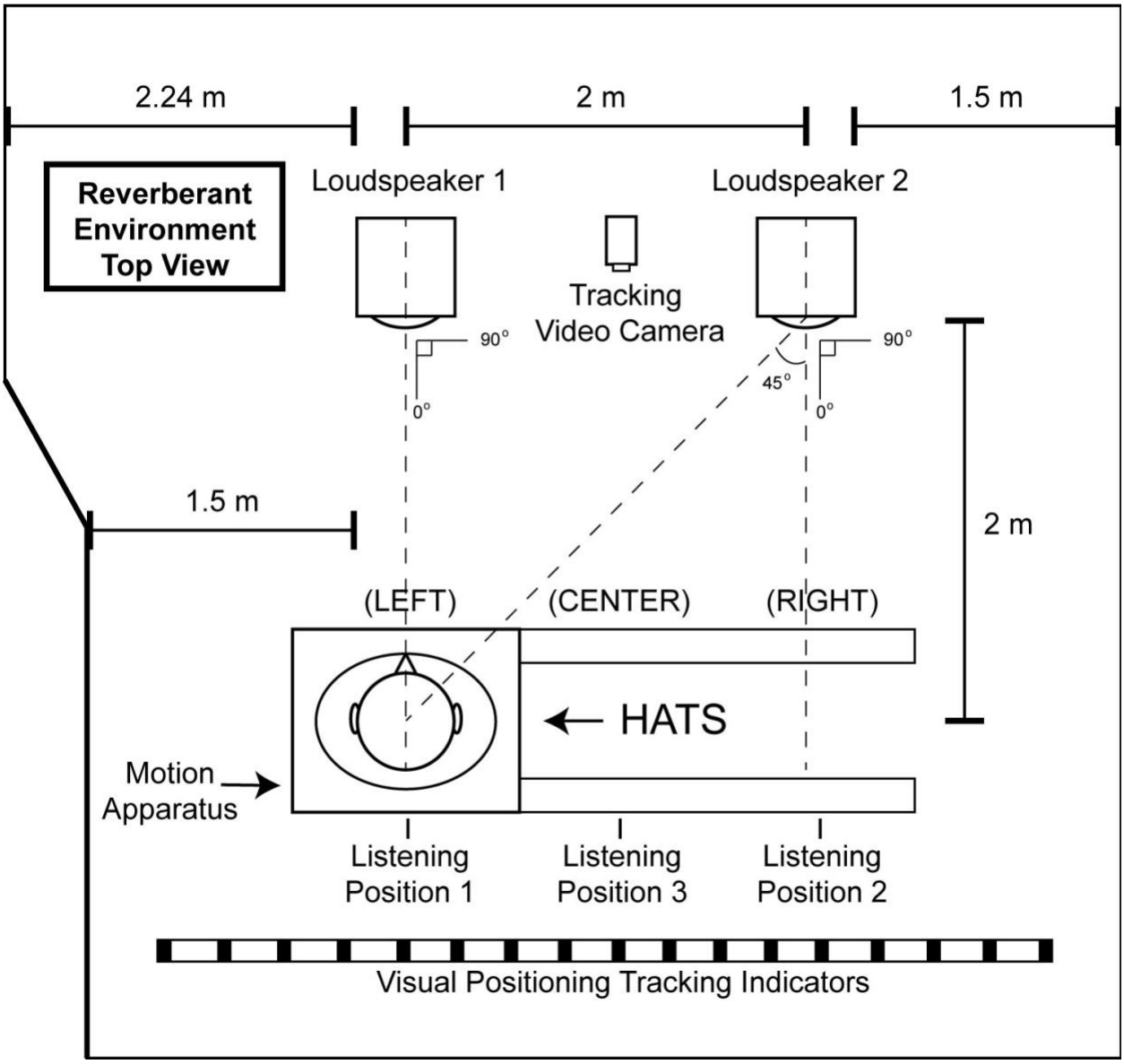

Figure 49: Diagram of the reverberant environment setup. Figure not drawn to scale.

BIBLIOGRAPHY

[1] O. Schreer et al., Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media, London: Wiley & Sons, 2013.

[2] R. Oldfield et al., "An object-based audio system for interactive broadcasting," in *AES 137th Convention*, Los Angeles, 2014.

[3] A. Wisbey, "How Streaming Object Based Audio Might Work," in *AES 143rd Convention*, New York, 2017.

[4] H. Fastl and E. Zwicker, Pyscho-Acoustics: Facts and Models, 3rd ed., New York: Springer, 2007.

[5] S. Paul, "Binaural Recording Technology: A Historical Review and Possible Future Developments," *Acustica United with Acustica,* vol. 95, pp. 767-788, 2009.

[6] J. W. Strutt and B. Rayleigh, The Theory of Sound, London: MacMillan and Co., 1877.

[7] J. Blauert, *Untersuchungen zum Richtungshören in der Medianebene bei fixiertem Kopf. (Investigations on directional hearing in the median plane with a fixed head.),* Achen: Rheinisch-Westfälische Technische Hochschule, 1969.

[8] J. Blauert, Spatial Hearing, London: The MIT Press, 1983.

[9] G. C. Stecker and F. J. Gallun, "Binaural Hearing, Sound Localization and Spatial Hearing," in *Translational Perspectives in Auditory Neuroscience: Normal Aspects of Hearing*, San Diego, Plural Publishing, 2012, pp. 383-483.

[10] B. Scharf, "Critical bands," in *Foundations of Modern Auditory Theory*, J. V. Tobias, Ed., New York, Academic Press, 1970, pp. 157-202.

[11] A. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound, Cambridge: MIT Press, 1990.

[12] J. Eargle, Handbook of Recording Engineering, New York: Van Nostrand Reinhold Company Inc., 1986.

[13] J. Braasch, "Modelling of Binaural Hearing," in *Communication Acoustics*, J. Blauert, Ed., Springer, 2005, pp. 76-108.

[14] J. Blauert, "Spatial Hearing with One Sound Source," in *Spatial Hearing: The*

Psychophysics of Human Sound Localization, London, The MIT Press, 1983, pp. 37-50, 63-70.

[15] L. Jeffress and R. W. Taylor, "Lateralization vs Localization," *The Journal of the Acoustical Society of America,* vol. 33, p. 482, 1961.

[16] C. C. Pratt, "The Spatial Character of High and Low Tones," *Journal of Experimental Psychology,* vol. 13, pp. 278-285, 1930.

[17] S. K. Roffler and R. A. Butler, "Factors that Influence the Localization of Sound in the Vertical Plane," *Journal of the Acoustical Society of America,* vol. 43, pp. 1255-1259, 1968.

[18] S. K. Roffler and R. A. Butler, "Localization of Tonal Stimuli in the Vertical Plane," *Journal of the Acoustical Society of America,* vol. 43, pp. 1260-1266, 1968.

[19] J. Blauert, "Sound Localization in the Median Plane," *Acustica,* vol. 22, pp. 205-213, 1970.

[20] R. Butler and R. Humanski, "Localization of Sound in the Vertical Plane with and without High Frequency Spectral Cues," *Perception & Psychophysics,* vol. 51, no. 2, pp. 182-186, 1992.

[21] J. Hebrank and D. Wright, "Spectral Cues Used in the Localization of Sound Sources on the Median Plane," *Journal of the Acoustical Society of America,* vol. 56, pp. 1829-1834,

1974.

[22] E. Georganti et al., "Extracting Sound-Source-Distance Information from Binaural Signals," in *The Technology of Binaural Listening*, J. Blauert, Ed., New York, Springer, 2013, pp. 174-178.

[23] W. King and D. Laird, "The Effect of Noise Intensity and Pattern on Locating Sounds," *Journal of the Acoustical Society,* pp. 99-102, 1930.

[24] A. Mills, "On the Minimum Audible Angle," *The Journal of the Acoustical Society of America,* vol. 30, no. 4, pp. 237-246, 1958.

[25] J. D. Harris and R. L. Sergeant, "Monaural/Binaural Minimum Audible Angles for a Moving Sound Source," *Journal of Speech and Hearing Research,* vol. 14, pp. 618-629, 1971.

[26] D. W. Grantham et al., "Auditory spatial resolution in horizontal, vertical, and diagonal planes," *The Journal of the Acoustical Society of America,* vol. 114, no. 2, pp. 1009-1022, 2003.

[27] W. O. Brimijoin and M. A. Akeroyd, "The Moving Minimum Audible Angle is Smaller During Self Motion than During Source Motion," *Frontiers in Neuroscience,* vol. 8, p. 273,

2014.

[28] S. Carlile and J. Leung, "The Perception of Auditory Motion," vol. 20, pp. 1-19, 2016.

[29] H. Wallach, E. B. Newman and M. R. Rosenzweig, "The Precedence Effect in Sound Localization," *The American Journal of Pyschology,* vol. 62, no. 3, pp. 315-336, 1949.

[30] J. P. Lochner and J. F. Burger, "The Subjective Masking of Short Time Delayed Echoes, Their Primary Sounds, and Their Contribution to the Intelligibility of Speech," *Acustica,* vol. 8, pp. 1-10, 1958.

[31] H. Haas, "The Influence of a Single Echo on the Audibility of Speech," *Journal of the Audio Engineering Society,* vol. 20, no. 2, pp. 146-159, 1972.

[32] D. Begault, "Virtual Acoustic Applications," in *In 3-D Sound for Virtual Reality and Multimedia*, Hanover, NASA, 2000, pp. 155-198.

[33] J. Blauert, Ed., The Technology of Binarual Listening, New York: Springer, 2013.

[34] D. Hammershøi and H. Møller, "Binaural Technique," in *Communication Acoustics*, J. Blauert, Ed., Berlin, Springer, 2005, pp. 223-254.

[35] D. Hammershøi and H. Møller, "Sound Transmission to and within the Human Ear Canal,"

Journal of the Acoustical Society of America, vol. 100, pp. 408-427, 1996.

[36] B. Gardner and K. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone," MIT Media Lab, Cambridge, 1994.

[37] V. R. Algazi et al., "The Cipic HRTF Database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2001.

[38] IRCAM & AKG, "LISTEN HRTF Database," 2002. [Online]. Available: http://recherche.ircam.fr/equipes/salles/listen/index.html. [Accessed 7 December 2019].

[39] J. H. McClellan et al., DSP First, 2nd ed., Boston: Pearson, 2016.

[40] S. Müller and P. Massarani, "Transfer-Function Measurement Using Sweeps," *Journal of the Audio Engineering Society,* vol. 49, pp. 443-471, 2001.

[41] D. Begault, "Reverberation," in *3-D Sound for Virtual Reality and Multimedia*, NASA, 2000, pp. 82-85.

[42] J. Borish and J. B. Angell, "An Efficient Algorithm for Measuring the Impulse Response Using Pseudorandom Noise," *Journal of the Audio Engineering Society,* vol. 31, pp. 478-488, 1983.

[43] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in *108th Audio Engineering Society Convention*, Paris, 2000.

[44] S. Foster, "Impulse Response Measurement Using Golay Codes," in *International Conference on Acoustics, Speech, and Signal Processing*, New York, 1986.

[45] C. I. Cheng and G. H. Wakefield, "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," *Journal of the Audio Engineering Society,* vol. 49, no. 4, pp. 231-249, 2001.

[46] R. E. Jensen and T. S. Welti, "The Importance of Reflections in a Binaural Room Impulse Response," in *114th Audio Engineering Society Convention*, Amsterdam, 2003.

[47] P. Zahorik, "Perceptually Relevant Parameters for Virtual Listening Simulation of Small Room Acoustics," *Journal of the Acoustical Society of America,* vol. 126, no. 2, pp. 776-791, 2009.

[48] T. Welti and S. Olive, "Validation of the Binarual Room Scanning Method Using Subjective Ratings of Spatial Attributes," in *48th International Audio Engineering Conferenece*, Munich, 2012.

[49] T. Z. Strybel et al., "Auditory Apparent Motion Under Binaural and Monaural Listening

Conditions," *Perception & Psychophysics,* vol. 45, no. 4, pp. 371-377, 1989.

[50] T. Z. Strybel and N. W., "The effect of burst duration, interstimulus onset interval, and loudspeaker arrangement on auditory apparent motion in the free field," *The Journal of the Acoustical Society of America,* vol. 96, no. 6, pp. 3463-3475, 1994.

[51] D. Phillips and S. Hall, "Spatial and temporal factors in auditory saltation," *The Journal of the Acoustical Society of America,* vol. 110, no. 3, pp. 1539-1547, 2001.

[52] R. Briggs and D. R. Perrott, "Auditory Apparent Movement Under Dichotic Listening Conditions," *Journal of Experimental Pyschology,* vol. 92, pp. 83-91, 1972.

[53] H. E. Burtt, "Auditory Illusions of Movement-A Preliminary Study," *Journal of Experimental Psychology,* vol. 2, pp. 63-75, 1917.

[54] O. Klemm, "Untersuchungen über die Lokalisation von Schallreizen. 3. Mitteilung: Über der Anteil des beidohrigen horens [Studies on the localization of sound stimuli. 3rd section: About the proportion of both ears]," *Archiv für die gesamte Psychologie,* vol. 38, pp. 71-114, 1918.

[55] A. Mathiesen, "Apparent Movement in Auditory Perception," *Pyschological Monographs,* vol. 41, no. 4, pp. 74-131, 1931.

[56] E. G. Boring, Psychology, Sensation and Perception in the History of Experimental, New York: Appleton-Century, 1942.

[57] V. Urbantschitsch, On disturbances of the equilibrium and illusory motions, 1897: Z. Ohrenheilkd, 1897.

[58] B. Riecke et al., "Moving Sounds Enhance the Visually-Induced Self-Motion Illusion (Circular Vection) in Virtual Reality," *ACM Transactions on Applied Perception,* vol. 6, no. 2, pp. 7:1 - 7:27, 2009.

[59] F. J. Calabro et al., "Acoustic Facilitation of Object Movement Detection During Self-Motion," *Proceedings of The Royal Society,* vol. 278, pp. 2840-2847, 2011.

[60] M. Gerzon, "Periphony: With-Height Sound Reproduction," *Journal of the AUdio Engineering Society,* vol. 21, no. 1, pp. 2-10, 1973.

[61] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Journal of the Audio Engineering Society,* vol. 45, no. 6, pp. 456-466, 1997.

[62] D. G. Malham and A. Myatt, "3-D Sound Spatialization Using Ambisonic Techniques," *Computer Music Journal,* vol. 19, no. 4, pp. 58-70, 1995.

[63] I. Frissen et. al, "Auditory Velocity Discrimination in the Horizontal Plane at Very High

Velocities," *Hearing Research,* vol. 316, pp. 94-101, 2014.

[64] N. Sankaran and S. Carlile, "Effects of Virtual Speaker Density and Room Reverberation on Spatiotemporal Thresholds of Audio-Visual Motion," *PloSone,* vol. 9, no. 9, 2014.

[65] B. Riecke et al., "Auditory Self-Motion Illusions ("Circular Vection") can be Facilitated by Vibrations and the Potential for Actual Motion," *APGV,* pp. 147-154, 2008.

[66] A. Nykänen et al., "Effects on Localization Performance from Moving the Sources in Binaural Reproductions," in *Inter Noise*, Innsbruck, Austria, 2013.

[67] P. Hoffmann and H. Møeller, "Audibility of Time Switching in Dynamic Binaural Synthesis," in *AES 118th Convention*, Barcelona, 2005.

[68] F. Freeland et al., "Efficient HRTF Interpolation in 3D Moving Sound," in *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, 2002.

[69] L. Savioja et al., "Creating Interactive Virtual Acoustic Environments," *Journal of the Audio Engineering Society,* vol. 47, no. 9, pp. 675-705, 1999.

[70] P. Minnaar et al., "Directional Resolution of Head-Related Transfer Functions Required in Binaural Synthesis," *Journal of the Audio Engineering Society,* vol. 53, no. 10, pp. 919-

929, 2005.

[71] U. Sloma et al., "Synthesis of Binaural Room Impulse Responses for Different Listening Positions Considering the Source Directivity," in *Audio Engineering Society 147th Convention*, New York, 2019.

[72] G. Enzner et al., "Trends in Acquisition of Individual Head-Related Transfer Functions," in *The Technology of Binaural Listening*, J. Blauert, Ed., Berlin, Springer, 2013, pp. 57-92.

[73] M. Geier et al., "The soundscape renderer: A versatile framework for spatial audio reproduction," in *Proceedings of the DEGA WFS Symposium*, Ilmenau, Germany, 2007.

[74] M. Vorländer, Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality, Berlin: Springer, 2008.

[75] M. Matsumoto and M. T., "Algorithms for Moving Sound Images," in *AES 114th Convention*, Amsterdam, 2003.

[76] A. Farnell, Designing Sound, Cambridge, MA: The MIT Press, 2010.

[77] International Telecom Union, "ITU-R BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," ITU, Geneva, 2015.

[78] International Telecom Union, "ITU-T P.800: Methods for Subjective Determination of Transmission Quality," ITU, Geneva, 1996.

[79] International Telecom Union, "ITU-R BS.1116-3: Methods for the Subjective Assessment of Small Impairments in Audio Systems," ITU, Geneva, 2015.

[80] A. Väljamäe, "Auditorily-Induced Illusory Self-Motion: A Review," *Brain Research Reviews,* vol. 61, pp. 240-255, 2009.

[81] M. Narbutt et al., "Streaming VR for Immersion: Quality Aspects of Compressed Spatial Audio," in *23rd International Conference on Virtual Systems & Multimedia*, Toronto, 2017.

[82] F. Rumsey, "Sound Quality," in *Sound and Recording*, 6th ed., Burlington, MA, Elsevier, 2009, pp. 563-590.

[83] S. Olive, "A Method for Training Listeners and Selecting Program Material for Listening Tests," in *Audio Engineering Society 97th Convention*, San Francisco, 1994.

[84] S. Olive, "Differences in Performance and Preference of Trained versus Untrained Listeners in Loudspeaker Tests: A Case Study," *Journal of the Audio Engineering Society,* vol. 51, no. 9, pp. 806-825, 2003.

[85] F. Rumsey et al., "Relationships Between Experienced Listener Ratings of Multichannel

Audio Quality and Naïve Listener Preferences," *The Journal of the Acoustical Society of America,* vol. 117, no. 6, pp. 3832-3840, 2005.

[86] S. Bech, "Selection and Training of Subjects for Listening Tests on Sound-Reproducing Equipment," *Journal of the Audio Engineering Society,* vol. 40, no. 7/8, pp. 590-610, 1992.

[87] W. Howie et al., "Effect of Audio Production Experience, Musical Training, and Age on Listener Performance in 3D Audio Evaluation," *Journal of the Audio Engineering Society,* vol. 67, no. 10, pp. 782-794, 2019.

[88] A. Kohlrausch and S. van de Par, "Audio–Visual Interaction in the Context of Multi-Media Applications," in *Communication Acoustics*, J. Blauert, Ed., Bochum, Springer, 2005, pp. 109-138.

[89] F. C. Fortenbaugh et al., "Sustained Attention Across the Life Span in a Sample of 10,000: Dissociating Ability and Strategy," *Psychological Science,* vol. 26, no. 9, pp. 1497-1510, 2015.

[90] M. S. Dobreva et al., "Influence of Aging on Human Sound Localization," *Journal of Nuerophysiology,* vol. 105, no. 5, pp. 2471-2486, 2011.

[91] S. Bech and N. Zacharov, "Test Planning, Administration and Reporting," in *Perceptual Audio Evaluation--Theory , Method and Application*, West Sussex, England, John Wiley &

Sons, 2006, pp. 301-321.

[92] T. Holman, "Mixing," in *Sound for Film and television*, 3rd ed., Burlington, MA: Focal Press, 2010, p. 185.

[93] "The Telephone at the Paris Opera," *Scientific American,* pp. 422-423, 31 December 1881.

[94] A. D. Blumlein, "Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems". United Kingdom Patent GB394325 (A), 14 June 1931.

[95] International Telecom Union, "ITU-R BS.775-3: Multichannel Stereophonic Sound System with and without Accompanying Picture," ITU, Geneva, 2012.

[96] International Telecom Union, "ITU-R BS.2159-7: Multichannel Sound Technology in Home and Broadcasting Applications," ITU, Geneva, 2015.

[97] F. Rumsey and T. McCormick, Sound and Recording: Applications and Theory, 7th ed., Burlington, MA: Focal Press, 2014.

[98] J. Hull, "Surround Sound," in *Handbook for Sound Engineers*, 5th ed., Burlington, MA: Focal Press, 2015.

[99] NHK Science & Technology Research Laboratory, "22.2 Multichannel Audio Format

Standardization Activity," *Broadcast Technology,* vol. 45, pp. 14-19, 2011.

[100] Dolby, "Dolby Atmos Specifications," Dolby, 2015.

[101] A. Berkhout, "A Holographic Approach to Acoustic Control," *Journal of the Audio Engineering Society,* vol. 36, pp. 977-995, 1988.

[102] R. King, Recording Orchestra and Other Classical Music Ensembles, New York: Routledge, 2017.

[103] T. Funkhouser et al., "Survey of Methods for Modeling Sound Propogation in Interactive Virtual Environment Systems," pp. 1-53, 2003.

[104] B. Karlheinz et al., "Wave Field Synthesis," in *3DTV Conference*, Potsdam, Germany, 2009.

[105] M. Chion and C. Gorbman, "Audio-Vision: Sound on Screen," New York, Columbia University Press, 1994.

[106] I. P. Howard and W. B. Templeton, Human Spatial Orientation, New York: Wiley, 1966.

[107] W. R. G. Thurlow and C. E. Jack, "Certain Determinants of the “Ventriloquist Effect”," *Percept Motor Skills,* vol. 36, no. 3, pp. 1171-1184, 1973.

[108] B. Sekular and R. Sekuler, "Collision Between Moving Visual targets: What Controls Alternative Ways of Seeing an Ambiguous Display," *Perception,* vol. 28, pp. 415-432, 1999.

[109] K. Watanabe and S. Shimojo, *Doctoral Dissertation: Crossmodal Interaction in Humans,* Pasadena: California Institute of Technology, 2000.

[110] J. Blauert, "Analysis and Synthesis of Auditory Scenes," in *Communication Acoustics*, New York, Springer-Verlag Berlin Heidelberg, 2005, pp. 12-13, 76-80.

[111] R. Algazi et al., "Motion-Tracked Binaural Sound," in *AES 116th Convention*, Berlin, 2004.

[112] B. Kapralos et al., "Auditory Cues in the Perception of Self Motion," in *AES 116th Convention*, Berlin, 2004.

[113] S. Bech and N. Zacharov, "Quantification of Impression," in *Perceptual Audio Evaluation--Theory, Method and Application*, West Sussex, England, John Wiley & Sons, 2006, pp. 39-96.

[114] J. MacDonald and P. Tran, "Loudspeaker Equalization for Auditory Research," in *Behavior Research Methods*, Springer-Verlag, 2007, pp. 133-136.

[115] International Telecom Union, "ITU-R BS.1770-4: Algorithms to Measure Audio

Programme Loudness and True-Peak Audio Level," ITU, Geneva, 2012.

[116] S. Bech and N. Zacharov, "Experimental Variables: Signal," in *Perceptual Audio Evaluation--Theory, Method and Application*, West Sussex, England, John Wiley & Sons Ltd, 2006, pp. 99-102.

[117] J. Neuhoff, "Auditory Motion and Localization," in *Ecological Pyschology*, Elsevier Academic Press, 2004, pp. 89-106.

[118] F. Brinkmann, *Doctoral Dissertation: Binaural Processing for the Evaluation of Acoustical Environments,* Berlin: Technische Universität Berlin, 2019.

[119] G.-B. Stan, "On the Use of Impulse Responses in High-End 3D Audio Virtual Reality Systems," Imperial College London, London.

[120] C. Tsakostas and A. Floros, "Real-time Spatial Representation of Moving Sound Sources," in *AES 123rd Convention*, New York, 2007.

[121] International Telecom Union, "ITU-T P.831: Subjective Performance Evaluation of Network Echo Cancellers," ITU, Geneva, 1998.

[122] International Telecom Union, "ITU-T P.913: Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution

Quality Television in Any Environment," ITU, Geneva, 2016.

[123] E. Wenzel and S. Foster, "Perceptual Consequences of Interpolating Head-Related Transfer Functions During the Spatial Synthesis," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Platz, NY, 1993.

[124] S. Granqvist, "The visual sort and rate method for perceptual evaluation in listening tests," *Logopedics Phoniatrics Vocology,* vol. 28, no. 3, pp. 109-116, 2003.

[125] M. Boerum et al., "Lateral Listener Movement on the Horizontal Plane: Sensing Motion Through Binaural Simulation," in *AES 61st International Conference: Audio for Games*, London, 2016.

[126] M. Mehta et al., "Design of Rooms for Speech," in *Architectural Acoustics*, London, Prentice Hall, 1999, p. 209.

[127] M. Vorländer, "Auralization of Spaces," *Physics Today,* vol. 62, no. 6, pp. 35-40, 2009.

[128] F. A. Everest and K. C. Pohlmann, "Acoustics of Listening Rooms," in *Master Handbook of Acoustics*, 5th ed., New York, McGraw Hill, 2009, p. 338.

[129] A. Nykänen et al., "Effects on Localization Performance from Moving the Sources in Binaural Reproductions," in *INTERNOISE and NOISE-CON Congress and Conference*

Preceedings, 2013.

[130] S. Devore et al., "Accurate Sound Localization in Reverberant Environments Is Mediated by Robust Encoding of Spatial Cues in the Auditory Midbrain," *Neuron,* vol. 62, pp. 123-134, 2009.

[131] W. M. Hartmann, "Listening in a Room and the Precedence Effect," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds., New York, Psychology Press, 2014, p. 194.

[132] G. Kearney et al., "Auditory Distance Perception with Static and Dynamic Binaural Rendering," in *AES 57th International Conference*, Hollywood, 2015.

[133] M. Rychtarikova et al., "Binaural Sound Source Localization in Real and Virtual Rooms," *Journal of the Audio Engineering Society,* vol. 57, no. 4, pp. 205-220, 2009.

[134] F. Chen, "Localization of 3-D Sound Presented through Headphone—Duration of Sound Presentation and Localization Accuracy," *Journal of the Audio Engineering Society,* vol. 51, no. 12, pp. 1163-1171, 2003.

[135] V. R. Algazi and R. Duda, "Approximating the Head-Related Transfer Function Using Simple Geometric Models of the Head and Torso," *Journal of the Acoustical Society of*

America, vol. 105, no. 5, pp. 2053-2064, 2002.

[136] Unity, "Unity-Manual: Reverb Zones," 2016. [Online]. Available: https://docs.unity3d.com/Manual/class- AudioReverbZone.html. [Accessed 26 08 2016].

[137] D. Tan et al., "Physically Large Displays Improve Path Integration in 3D Virtual Navigation Tasks," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, 2004.

[138] E. Patrick et al., "Using a Large Projection Screen as an Alternative to Head-Mounted Displays for Virtual Environments," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, 2000.

[139] M. E. Grabe et al., "T.B. The role of screen size in viewer experiences of media content," *Visual Communication Quarterly,* vol. 6, pp. 4-9, 1999.

[140] F. Rumsey, "Spatial Audio: Binaural Challenges," *Journal of the Audio Engineering Society,* vol. 62, no. 11, 2014.

[141] AudioKinetic, "Wwise," 2018. [Online]. Available: https://www.audiokinetic.com/products/wwis e. [Accessed 1 5 2018].

[142] Google, "Tilt Brush," 2018. [Online]. Available: https://www.tiltbrush.com. [Accessed 1 5

2018].

[143] Audio Fusion, "Virtual Studio 3D," 2018. [Online]. Available: https://audiofusion.com/vas3d/. [Accessed 1 5 2018].

[144] J. Rose, "The Mix, Putting Things in Perspective," in *Audio Postproduction for Digital Video*, San Francisco, CMP Books, 2012, pp. 377-84.

[145] F. Rumsey, "Applications of Spatial Audio," in *Spatial Audio*, New York, Focal Press, 2001, pp. 18-19.

[146] J. Blauert, "Binaural Room Simulation and Auditory Virtual Reality, Binaural Signal Processing, the Precedence Effect," in *Spatial Hearing*, 2nd ed., Cambridge, The MIT Press, 1997, pp. 370-418.

[147] J. Udesen et al., "The Effect of Vision on Psychoacoustic Testing with Headphone- Based Virtual Sound," *Journal of the Audio Engineering Society,* vol. 63, no. 7/8, pp. 552-561, 2015.

[148] S.-W. Jeon et al., "Virtual Source Panning Using Multiple-Wise Vector Base in the Multispeaker Stereo Format," in *19th European Signal Processing Conference*, Barcelona, 2011.

[149] G. Marentakis et al., "Vector-Base and Ambisonic Amplitude Panning: A Comparison Using Pop, Classical, and Contemporary Spatial Music," *Acustica United with Acustica,* vol. 100, no. 5, pp. 945-955, 2014.

[150] J. M. Jot, "Spatialisateur," IRCAM, 2012.

[151] E. H. Langendijk and A. W. Bronkhorst, "Contribution of Spectral Cues to Human Sound Localization," *The Journal of the Acoustical Society of America,* vol. 112, no. 4, pp. 1583-1596, 2002.

[152] L. M. Kells, W. F. Kearn and J. R. Bland, Plane and Spherical Trigonometry, New York: McGraw-Hill, 1940.

[153] W. J. Conover, "Statistics of the Kolmogorov- Smirnov Type," in *Practical Nonparametric Statistics*, 3rd ed., New York, John Wiley & Sons, 1999.

[154] J. Utts and R. Heckard, "Analysis of Variance," in *Mind on Statistics*, Duxbury, 2002, pp. 496-506.

[155] A. Nowak-Brzeziński, "Outlier Mining in Rule-Based Knowledge Bases," in *8th International Conference, RSCTC 2012*, Berlin, 2012.

[156] D. Benson, "Minimum Audible Movement Angles for Discriminating Upward from

Downward Trajectories of Smooth Virtual Source Motion within a Sagittal Plane," McGill University, Montreal, Canada, 2007.

[157] J. Melick et al., "Customization for Personalized Rendering of Motion-Tracked Binaural Sound," in *117th Audio Engineering Society Convention*, San Francisco, 2004.

[158] S. E. Garcia et al., "Auditory Localisation Biases Increase with Sensory Uncertainty," *Scientific Reports,* vol. 7, no. 40567, 2017.

[159] R. Baumgartner et al., "Asymmetries in Behavioral and Neural Responses to Spectral Cues Demonstrate the Generality of Auditory looming Bias," *Proceedings of the National Academy of Sciences,* vol. 114, no. 36, pp. 9743-9748, 2017.

[160] J. C. Goodwin, "Control Problems in Experimental Research: Problems with Biasing," in *Research in Psychology: Methods and Design*, Wiley, 2009, pp. 232-235.