A Spatial Audio Remote Presence Experience for Musicians

Antoine Pierre Augustin Aubet

Department of Electrical and Computer Engineering
McGill University
Montreal, Canada

April 2025

A thesis submitted to McGill University in partial fulfilment of the requirements for the degree of Master of Science.

Abstract

One of the main goals of technology has been to reduce distances between people by making it easier and faster to communicate across great distances. Nowadays, one can video-call someone from across the Earth in high-resolution video and high-quality audio and have a near-instantaneous two-way conversation. The next logical step is to make people that are physically distant feel like they are in the same location, which is a concept is known as co-presence.

This thesis describes an attempt to create a sense of co-presence through sound, such that each participant perceives the others to be present in the room alongside themselves. A model of the room is created using its dimensions and approximate reflection coefficients for each surface. Using the incoming audio from the other participant, the model is used to place a virtual source in the listener's space and model the sound field where the listener is located using Ambisonics. By tracking the listener's head orientation, we can decode the sound field into binaural audio corresponding to their head orientation and play the other participant's sound through headphones into their ears the way they would perceive it if the other participants were physically present in the room.

We tested this approach on pairs of musicians since they have high standards for audio quality and fidelity that are more demanding than most other disciplines. We found that the Ambisonics decoding into binaural audio added 34.5 ms of latency. Though it did not hamper their playing, the additional latency was perceivable by some participants, particularly the ones that played rhythmic songs. Furthermore, while some participants enjoyed the challenge of playing without seeing their partner, others mentioned missing the visual channel that they usually rely on to communicate and coordinate their playing. However, from the interviews, 60% of musicians mentioned preferring the sound experience in the condition where the spatial effect with the room acoustics was activated.

Based on the results obtained in this experiment, it is clear that reducing the latency by using a different plug-in would be a substantial improvement, making the remote playing

experience even more compelling. Furthermore, providing a visual channel for the musicians to see each other would improve musical communication, addressing the most mentioned complaint about the setup.

Résumé

L'un des principaux objectifs de la technologie a toujours été de réduire les distances entre les personnes en facilitant et en accélérant la communication à travers de grandes distances. Aujourd'hui, il est possible de passer un appel vidéo en haute résolution avec une qualité audio élevée à quelqu'un à l'autre bout du monde et d'avoir une conversation bidirectionnelle quasi instantanée. L'étape logique suivante est de faire en sorte que des personnes physiquement éloignées aient l'impression d'être au même endroit, un concept connu sous le nom de co-présence.

Cette thèse décrit une tentative de création d'un sentiment de co-présence par le son, de manière à ce que chaque participant perçoive les autres comme étant présents dans la pièce à ses côtés. Un modèle de la pièce est créé à l'aide de ses dimensions et des coefficients de réflexion approximatifs de chaque surface. En utilisant l'audio entrant de l'autre participant, le modèle permet de placer une source sonore virtuelle dans l'espace de l'auditeur et de modéliser le champ sonore à l'endroit où il se trouve grâce à l'Ambisonie. En suivant l'orientation de la tête de l'auditeur, nous pouvons décoder le champ sonore en audio binaural correspondant à son orientation et diffuser le son de l'autre participant dans ses écouteurs, de la manière dont il le percevrait si l'autre personne était physiquement présente dans la pièce.

Nous avons testé cette approche avec des duos de musiciens, car ils ont des exigences élevées en matière de qualité et de fidélité audio, plus strictes que dans la plupart des autres disciplines. Nous avons constaté que le décodage de l'Ambisonie en audio binaural ajoutait une latence de 34.5 ms. Bien que cela n'ait pas entravé leur jeu, certains participants, en particulier ceux qui jouaient des morceaux rythmiques, ont perçu cette latence. De plus, tandis que certains musiciens ont apprécié le défi de jouer sans voir leur partenaire, d'autres ont mentionné le manque du canal visuel sur lequel ils s'appuient habituellement pour communiquer et se coordonner. Toutefois, d'après les entretiens, 60% des musiciens ont déclaré préférer l'expérience sonore lorsque l'effet spatial avec l'acoustique de la pièce

était activé.

À la lumière des résultats obtenus dans cette expérience, il est clair que la réduction de la latence grâce à l'utilisation d'un autre plug-in constituerait une amélioration significative, rendant l'expérience de jeu à distance encore plus convaincante. De plus, fournir un canal visuel permettant aux musiciens de se voir améliorerait la communication musicale, répondant ainsi à la critique la plus fréquemment évoquée concernant cette configuration.

Acknowledgments

First of all, I would like to thank Professor Jeremy Cooperstock for his guidance and mentorship. He allowed me freedom to explore various topics and projects while providing me with the resources to succeed in my research journey. I would not have walked the same path without his connections or valuable directions.

I would also like to thank the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) for their support. In particular, Julien Boissinot for sharing his expert knowledge through insightful conversations and showing me the way to better solutions, and to Yves Méthot for his technical advice and support. I would also like to thank Professor Marcelo Wanderley for his encouragement, positivity, and his different perspectives on my project.

Through this adventure, my colleagues at the Shared Reality Lab have also been incredibly resourceful. Specifically, Cyan Kuo for sharing her wisdom and advice, Juliette Regimbal for her qualitative analysis techniques, Emmanuel Wilson for his willingness to make time for productive discussions, and Max Henry for his music technology knowledge giving rise to ever so useful suggestions, as well as for taking the time to review this document.

Last but certainly not least, I would like to thank Sayaka for her continuous encouragement and support during these two and a half years, pushing me to persevere when making progress was difficult.

Contents

List of Figures

List of Tables

List of Acronyms

ASIO	Audio Stream Input/Output
CD	compact disc
CIRMMT	Centre for Interdisciplinary Research in Music Media and Technology
CPU	central processing unit
DAW	digital audio workstation
DFT	discrete Fourier transform
DoF	degree of freedom
EDT	Early decay time
FFT	fast Fourier transform
FIR	finite impulse response
HCI	Human-Computer Interaction
HOA	higher order Ambisonics
HRIR	head-related impulse response
HRTF	head-related transfer function

IFFT	inverse fast Fourier transform
ILD	interaural level difference
IMU	inertial measurement unit
IRS	inverse repeated sequence
ITD	interaural time difference
JND	just-noticeable difference
KEMAR	Knowles Electronics Manikin for Acoustics Research
MLS	maximum-length sequence
OLA	overlap-add
OLS	overlap-save
OSC	Open Sound Control
REB	Research Ethics Board
RIR	room impulse response
RT	reverberation time
TCP	Transmission Control Protocol
UDP	User Datagram Protocol

Chapter 1

Introduction

1.1 Telepresence

Telepresence, also known as remote presence, is the concept of enabling a person to *feel* as if they are present, or to make it *seem like* they are present, at a location other than their true physical location, through the use of technology [1, 2]. It generally involves the use of communication and sensory technologies to transmit audio, video, or other sensory data between two physically distinct locations in real-time. This enables a sense of immersion and engagement that simulates physical presence, which can be beneficial in communication, events, or remote collaboration applications where a user feels as if they were participating in person. Under the Human-Computer Interaction (HCI) umbrella, it is an entire field of study where the research focuses on finding what factors humans require to perceive being physically present, how to reproduce these sensations, and how to make up for the shortcomings of certain senses in a remote setting. In a situation where telepresence is used to connect with other people, it can be called co-presence, referring to the shared presence of all participants despite not sharing the same physical space.

The sense of hearing greatly affects our perception of the environment and presence, so solutions need to be found to emulate the same psychoacoustic cues that are found in a given environment. One important factor is the sound quality. This is the assessment of the auditory image where the listener can either express their satisfaction or dissatisfaction with that image [3]. This quality is made up of two subqualities: *spatial* and *timbral* quality.

The former describes the distribution of the sound sources and the size, shape, and material of the space in which the sound is heard. For instance, a listener would be able to hear two different sounds that have the same pitch, loudness, duration, and timbre, and notice they are arriving from two separate locations.

The latter attribute, timbre, describes the spectral character of sound. An example that highlights the importance of this attribute is calling or conferencing software. Most of them compress the incoming audio to make it easier to transport over a network [4]. This means that certain frequencies will be removed as they are not deemed *critical* or *important* to the application since the purpose is only to be understood. Consequently, the voice is not perfectly reproduced at the other end of the line. The timbre of the voice is affected in a way that is not expected by the listener, meaning it does not sound very natural. The phone or computer speaker will also affect the sound, potentially making it sound 'tinny', meaning it cannot reproduce the lower frequencies accurately. Typically, telephone calls only transmit audio in the 300-3000 Hz frequency range. This approach was chosen as a compromise between speech intelligibility, retaining 80% of speech information, and bandwidth requirements for telephone networks. However, anything outside of this range is often discarded, including the harmonic frequencies that are part of the sound we make and add to the sound signature of a voice [5]. The removal of harmonics and possibly of the fundamental results in a sound that lacks naturalness. Although speech remains intelligible, the removal of higher frequencies in speech diminishes the intelligibility by approximately 20% [6].

During a basic telephone conversation, participants do not experience a sense of presence with their interlocutor, nor do they anticipate such an experience. By contrast, within virtual environments such as simulators, users expect these tools to replicate the sensations of real-world interactions, as it is their purpose, so they will be made to meet those expectations. In such cases, we cannot take shortcuts the way telephone sampling does. Of course, depending on the application, every sense might not be as important. In a welding simulator, the haptics might need to be the most realistic, while in a vehicle simulator, visual fidelity might be the most important. This is because for a student to practice welding, they need to get used to the feeling of it so that they can tune their movements, while someone learning to drive a car needs to be aware of their surroundings and pay close attention to what they see. The topic of this thesis is the creation of a telepresence system for musicians. It then follows that the most important sense for this application is hearing.

1.2 Sound

Listeners develop expectations regarding the characteristics of sounds generated by the other individual, whether through vocalizations, bodily actions (e.g., sneezing, clapping), or interactions with the environment (e.g., moving a chair). These expectations encompass how such sounds will be influenced by the physical setting in which they are produced [7]. This is called the soundscape [8]. That means that in a large, enclosed, untreated space, we would expect to hear echo as a result of the other person speaking. Conversely, in a small, carpeted space, there should be very little reverb coming from the space when the other person makes a sound. When the audio cues do not match the visual experience of the user, the incongruity between the senses can break the immersion that we are trying to create [9]. In a video call with perfect sound quality, if the setting for one participant is very different from the other, the acoustic properties for each space will likely be quite distinct. As a result, it will be clear to both parties that the other participant is in a separate space because the sound produced by their interlocutor does not match what they would expect coming from their own physical space. These shortcomings will be addressed in the work presented here.

1.3 Telepresence for Musicians

This thesis proposes a design concept for a two-participant telepresence setup, and then investigates its performance through a user study with musicians. The aim is to acoustically simulate the presence of another musician in one's space. This is so that each musician feels like they are in their room and the other is present with them. This approach simplifies the audiovisual congruence problem: by embracing each musician's space, their perception of their own respective environment does not need to be altered. Instead, the virtual components—the distant musician: their appearance and their sound—need to be conditioned to match the real environment. Using a virtual environment would not only require performing this alteration for the distant musician so that they match the virtual environment, but also require making the local musician's sound match the new environment and changing their visual perception of themselves. This would involve cutting out the musician's instrument and body from the visual of the new environment as well as changing the lighting and color of those to match the new environment since musicians want to see themselves and their own instrument, and seeing them be dark in a potentially bright virtual environment would result in incongruence. Therefore, we embrace the current environment they are present in and make the external musicians match that space. How to bring in the other musicians visually is still a question, though it will not be the subject of this thesis and remains in the future directions. Instead, we are interested in how to make it sound like, from one musician's perspective, another musician is present in the same room.

In the realm of telepresence, especially concerning networked music performance, musicians stand out as a group highly sensitive to latency, with ears finely attuned to even the subtlest characteristics and alterations in sound quality and features. Therefore, designing a system that aims to cater to them poses an interesting technical challenge to solve while also aiming to provide them with a satisfactory telepresence experience that they can use to practice and perform with minimal limitations.

This thesis focuses on the audio experience of the project, and design decisions were made as a consequence while keeping in mind the future integration of a visual feedback component. The technical details are laid out and explained in the Methods section and the design is put to the test through a user study. From our results, we can validate some design decisions such as the choice of the technology for room acoustics simulation; however, they also show the need for improvements in areas such as lowering the latency impact of the sound processing and enabling visual feedback between the musicians. A list of priorities is established to inform the further directions of this project and its eventual integration with other work in the visual sense.

Chapter 2

Background

2.1 Telepresence and Networked Music Performance

Telepresence for musicians is a concept that, in order to be achieved to its full extent, requires solving specific problems. As mentioned previously, musicians are very sensitive to latency, and as a result, any proposed system design should not add significant delay, with end-to-end latency not surpassing 30 ms where slow-downs can start occurring in the performance [10]. There will be a maximum physical distance between the participants as, even with data transmission at the speed of light, a certain distance would result in a latency that is unsuitable for playing at a certain tempo. The sound quality also needs to be high. Today, the standard is 24-bit, 92 kHz sampling. This is now a solved problem since domestic internet speeds often reach a gigabit per second, vastly sufficient even for multichannel, high-quality audio transmission. Finally, just like the word telepresence suggests, presence needs to be achieved. That is, how to convince the musician that what they are experiencing is real.

There have been multiple projects in the past attempting to solve some of the problems with telepresence for musicians. Ultra-Videoconferencing and JackTrip have demonstrated high-quality video combined with audio transmission for the former and high quality, low-latency audio transmission for the latter [11, 12, 13].

Ultra-Videoconferencing [11]

Ultra-Videoconferencing is a project from the Shared Reality Lab at McGill University. Its origins come from the desire to transmit high quality multichannel audio and video in real-time over the internet. In 1999, a musical performance took place at McGill University and was transmitted to an audience at New York University with a delay of approximately 3 s. The lower latency is achieved through multiple networking and processing tricks such as packet and buffer size optimization, the use of Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) in tandem to make use of the strength of each protocol, and utilizing dedicated, high throughput networks. Later iterations improved on these methods by using uncompressed audio and video to remove the stream compression delay and selecting hardware with smaller data buffers.

JackTrip [13]

JackTrip is an application developed by the SoundWire group at Stanford University. It is based on the ideas from Ultra-Videoconferencing but focuses on low-latency audio. It transmits uncompressed audio and minimizes latency for an n-directional musical performance all in one package. Design decisions such as using a circular buffer for inter-thread communication, real-time scheduling for threads, using UDP, and a packet redundancy algorithm all contribute to the low latency and stability of the tool.

2.1.1 Music Telepresence

Making someone feel as if they are in the presence of somebody else using technology is a challenge. In general, using the highest quality audio and video possible improves *immersion*. However, higher quality can lead to increased latency. In a video call, a lack of detail in the face of the interlocutor may break the immersion, just like a slow reaction time. For simple calls, current technology accomplishes high quality audio and video transmission with ease. However, for musical performance, the audio in particular is held to a higher standard, both in terms of quality and latency. To take immersion and engagement between participants further, we are now interested in implementing the room acoustics into the system.

The ultimate goal for this project is to build upon these projects and to provide a mixed reality experience where each musician sees and hears all the others in that musician's physical space. This means that, if a band were to play together with each member in a different location, each musician would perceive the others to be present in the room alongside themselves, as illustrated in Figure 2.1.

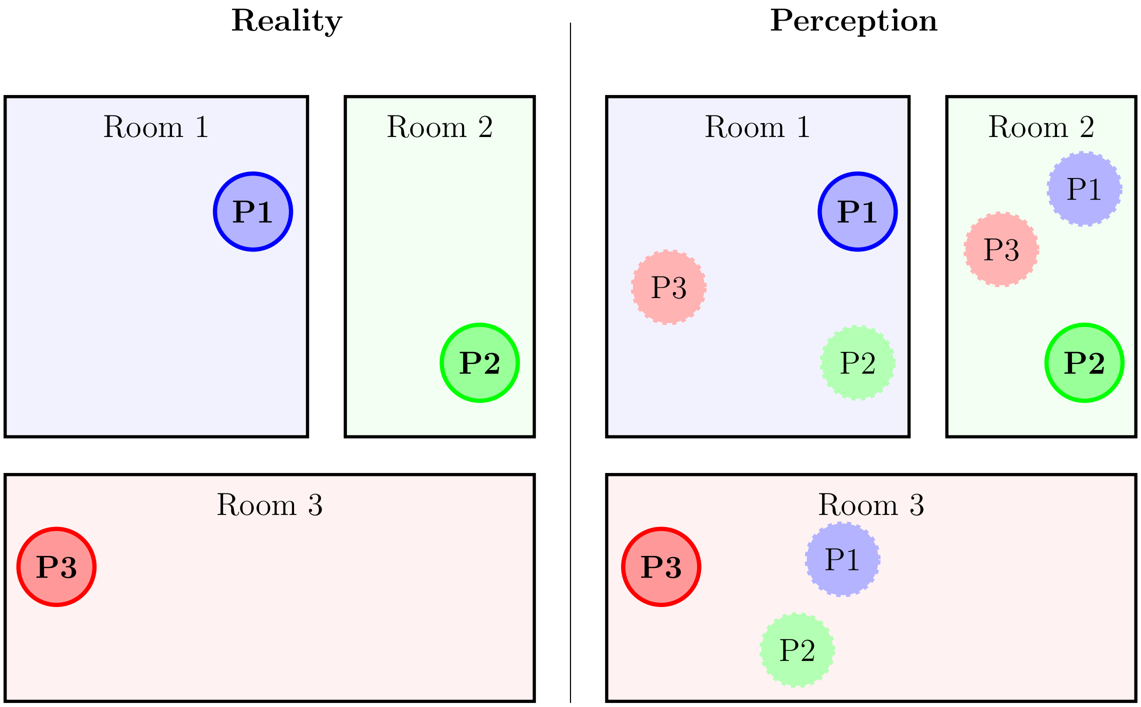

Figure 2.1 Illustration of the concept of the project. The left side of the figure illustrates the physical world where each participant (shown with an outlined circle) is in their own room. The right side illustrates how each participant perceives the other "simulated" participants (shown with an uneven edge) to be present with them. The simulated location of the virtual participants can be chosen by the musician depending on their preference, *i.e.*, closer or further away, central or in the corner, etc.

While both audio and video are important and each sense relies on the other to produce a compelling experience, they are separate pieces that can be worked on separately. This work focuses on the auditory channel in the system: what its components are, how they work, how they contribute to the experience, and finally its evaluation through a user study.

2.2 Sound and Perception

To create a good system that caters to the sense of hearing, we must first understand how sound is heard, so this section describes the mechanisms behind the way we perceive sound.

Sound is vibrations—or mechanical waves—propagating through a medium. The human auditory system is sensitive to vibration frequencies ranging between 20 and 20 000 Hz. The outer part of the ear, the pinna, is the most relevant in this research as it is the part of the ear with the greatest impact on directionality perception, which is discussed in the next section.

2.2.1 Directionality

When sound is produced by a source, the sound waves make their way to the ears. Depending on the direction the head is facing, the waves might hit both ears at once, or one ear before the other. This is known as interaural time difference (ITD). ITD is a major cue for humans when it comes to lateral audio spatialization, meaning how well humans can understand the direction a sound is coming from [14]. This effect is detectable starting with frequencies around 3000 Hz and down [15].

For frequencies above 1000 Hz, interaural level difference (ILD), which is the difference in sound level between both ears, also contributes to the human ability to determine horizontal positioning of a sound source [16].

The unique shape of each individual's head, shoulders, and torso influences sound perception by altering how sound waves reflect and bend around the body [17]. These variations mean that not everyone experiences sound in the same way.

Finally, the shape of the outer ear, or pinna, affects the way humans perceive sound [17]. Sound will bounce off of the ear's ridges, and each bounce impacts the sound waves. By the time the sound waves reach the ear canal, they will have been transformed, impacting each frequency intensity differently. Just as the shape of the upper body varies, so do each person's ears, further emphasizing that everyone perceives the same sound differently.

The just-noticeable difference (JND) for source location angle in humans is about 10° in most directions. For frontal sound sources, this can be as precise as 1° horizontally, depending on the spectral features of the sound [18]. Sources more than a meter away can have very similar localization cues if they are located anywhere on the surface of a cone centered on the interaural axis, a phenomenon known as the *cone of confusion* [19].

2.2.2 Distance

The human ability to estimate the distance of a sound source is much less accurate than the ability to determine the direction of that sound source [20]. While humans can detect changes in horizontal position within about 1°, large errors in distance estimation between 5% and 25% are common.

Listeners often overestimate the distance of nearby sources (less than 1 m) and underestimate the distance to faraway sources [20]. The most significant cue humans use to understand distance from the sound source is its intensity [21]. Under ideal conditions, sound intensity and distance are correlated by the inverse-square law, which states that for a point source radiating omnidirectionally, intensity decreases by 6 dB for every doubling of distance [22]. While it can still be a good acoustic cue for distance, humans are not very sensitive to sound intensity. Studies report that only around a 20% change in source distance is just noticeable [21, 23, 24], but some have found higher and lower thresholds, as low as 3% [25] and as high as 48% for close distances [26]. Other distance cues include:

- *spectrum*: some frequencies being absorbed by the transmission medium affect the perceived sound spectrum, mostly noticeable for distances greater than 15 m,

- *binaural cues*: through scattering effects of the head and torso,
- *dynamic cues*: cues related to motion of both the source and receiver
- *direct-to-reverberant energy ratio*

The direct-to-reverberant energy ratio is an important cue indoors since it relates the energy of the sound waves from the direct path and the energy from the surface reflections around the listener. These other distance cues are not as important and they have also been much less studied than the intensity. Non-acoustic cues, such as vision and familiarity, also play a part [20].

2.2.3 Effects of the Environment

The environment that a sound is produced in will have an effect on the way it is perceived. When a sound is produced, it will spread in all directions. Some of the sound wave will hit the receiver directly, while other parts of the wave will bounce off surfaces and objects and possibly also make it to the receiver following these reflections. This means that these paths will be longer than the direct path and, therefore, arrive at the receiver later than the sound waves taking the direct path. This is reverberation, colloquially known as *reverb*. The surfaces, shape, and volume of a room act as filters and will dampen certain frequencies more than others, which can be modeled by Green's function [27]. These sound waves are therefore modified and will sound different, both due to the filtering and the delay.

Particularly, outdoors, where there is very little reverb from surfaces, the direct path will have the highest intensity by far. Although sound intensity decreases by 6 dB outdoors for each doubling of distance, this reduction is only approximately 4 dB indoors due to the sound waves propagating in a closed environment [26]. As a result, reverberation has a significantly higher intensity indoors.

2.2.4 Spatial Audio

Spatial audio is the technology that enables a virtual source of sound to be fixed in space as the user turns their head around. This technology aims to replicate the way humans perceive sound in real life. This means, from a source signal, replicating the mechanisms that affect the sound waves in the same way that they would be affected by the environment such that the perception is the same to the listener. There are multiple ways to do this, which will be discussed in the subsequent sections.

Binaural Audio

Binaural audio is a sound recording technique that aims to capture sound the way our ears receive it. In sound processing, this means using a filter such that the input stream is transformed into what is received by the inner ear. This is achieved mathematically using a convolution operation in the time domain or a multiplication in the frequency domain.

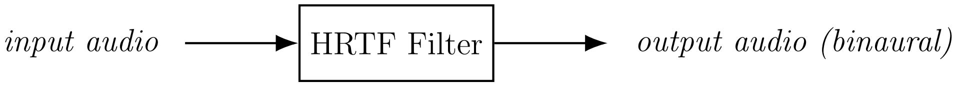

Figure 2.2 Block diagram of the transformation into binaural audio. The input is the sound signal from the environment, and the output is transformed the way the ear would perceive it.

Finding the right filter is the difficult part. Open-source databases exist for such filters, known as head-related transfer functions (HRTFs). These databases are created by playing back sounds through speakers and recording them using dummy ears, allowing the extraction of the transfer function from the input to the output [28]. By placing the microphone in a dummy ear, ideally as part of a dummy head and torso such as Knowles Electronics Manikin for Acoustics Research (KEMAR) [29], the head-related impulse response (HRIR) can be measured, from which we can obtain the HRTF. The filter changes according to the localization of the sound. Therefore, an entire database is needed for every angle possible. Usually, measurements are made every 1° to 5° on the horizontal plane and 5° to 10° on the vertical plane.

Using the appropriate HRTF for the source angle, a sound can theoretically be reproduced as if it were produced at that angle relative to the listener. Since everyone's perception is different due to the factors discussed in Section 2.2, we can only create an approximation or an average for everyone. The KEMAR dummy, which is considered to be the standard head and torso model for acoustic measurements [30], is based on the median measurement of about 5000 adult males [29]. The literature suggests that these approximations can sometimes not be good enough, and not work at all with some people. The KEMAR manikin is based on average adult measurements, making it unfit for children. Furthermore, it has also been shown to not be a great model for Asians due to the different shape of their heads [31]. Using an average model for HRTF is therefore not ideal, particularly when it works better with certain populations rather than others. While it is a complex task, it is always better to use personalized HRTFs since, as discussed above, everyone perceives sound in a unique way. However, more recently, after years of focus in research, progress was made on ways to personalize these HRTFs [17, 32].

A similar finite impulse response (FIR) measurement technique can be employed to create a filter that applies the acoustic characteristics of a specific environment. This method, analogous to the one described above, involves measuring the FIR, also known as the room impulse response (RIR) in the context of a room. The resulting filter is tailored to the source-receiver pair, accounting for their specific location and orientation within the setup. As a result, in theory, an array of source-receiver filters would be needed to cover an entire room if the source and listener will both move around the space.

To record a RIR, the signal being played back depends on the method used, with varying performance depending on the environment [33]. Some methods include:

- maximum-length sequence (MLS)
- inverse repeated sequence (IRS)

- time-stretched pulses
- SineSweep

The output will capture the effects of the various reflections and absorptions caused by the environment. From there, the difference between the two can give us the transfer function corresponding to the filter. This method can capture the effects of the environment that the FIR is measured in. It can also capture the effect of the ear shape as described above, such as with a KEMAR dummy. The downside of this method is that it captures an exact situation, meaning that if the orientation, location, or environment changes slightly, the filter is not *accurate* anymore. This makes this solution more suitable for static rather than dynamic environments.

Image-Source Model

Another method for simulating the environment is the image-source model, illustrated in Figure 2.3. This is a geometric simulation method that simulates the paths of sound reflections on surfaces, treating them like mirrors, between the source and the receiver. This method assumes that sound travels in straight lines, undergoing perfect reflection when encountering an obstacle, usually the walls, the ceiling, or the floor.

The challenge with this method lies in accurately determining the reflection coefficients (or functions) for each surface. Without precise reflection data, the model will fail to accurately replicate the real environment.

Sound Ray-Tracing

The sound ray-tracing method assumes straight-line propagation and simulates reflections off surfaces. In contrast to the image-source model, which is deterministic, the ray-tracing method is stochastic in nature.

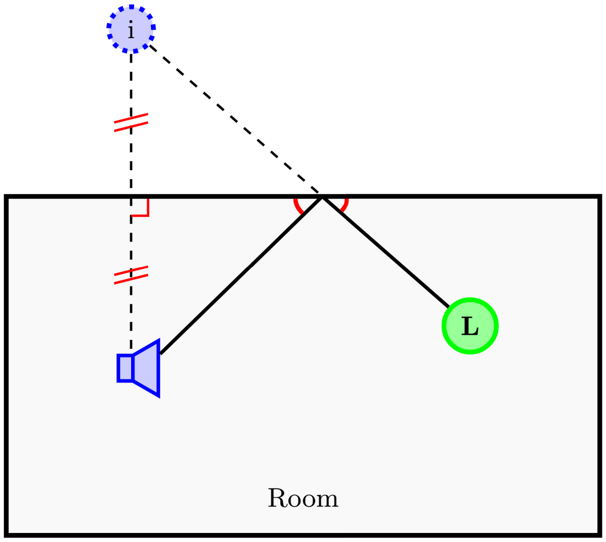

Figure 2.3 An illustration of the image-source model with a wall reflection. The image of the source is behind the wall forming a direct path from the image to the listener. Each intersection with a wall means applying the damping coefficient of that wall to the sound stream.

Ambisonics

Ambisonics is a surround sound format. Unlike typical sound formats such as stereo, Ambisonics does not attribute a channel for each speaker. Instead, it aims to capture a representation of the sound field that can then be decoded to be recreated by the listener's speaker setup. This means that the sound can be thought of as moving in a direction rather than coming from a certain speaker position, which is then easily adapted to the listener's layout and number of speakers. The way this is achieved is through spherical harmonics decomposition. The first channel is an omnidirectional channel, and each subsequent channel corresponds to a harmonic. For instance, in the first order, there are four channels: W, X, Y, and Z. The W channel is the omnidirectional channel, the X channel is the front-minus-back, the Y channel is left-minus-right, and Z is up-minus-down. This configuration arises because the first-order spherical harmonics generate figures-of-eight patterns along their respective axes. As the order increases, the accuracy of sound field reproduction improves; however, this necessitates the use of additional channels and more complex computational processes. A third-order Ambisonics signal contains 16 channels.

This format offers a practical approach for determining the sound field at the listener's location, where sound reflections originate from all directions. Ambisonics simplifies the task of tracking the sounds, their reflections, and their respective directions within a room, given its parameters for surface absorption. Subsequently, each reflection can be convolved with the HRTF corresponding to its direction and played back to the listener.

2.2.5 The Ventriloquist Effect

The ventriloquist effect is a perceptual phenomenon where the brain integrates visual and auditory information to determine the location of a sound source [34]. Named after the ventriloquists from ancient Greece and the Roman Empire, this effect demonstrates the brain's ability to make a voice appear to come from somewhere other than its actual source. This phenomenon occurs because human visual localization is typically more accurate than auditory localization. When the source of a sound appears to be visible and clear, the visual sense dominates in identifying the sound source. However, if the visual estimate is sufficiently compromised, the auditory sense can take over.

Chapter 3

System and User Study Design

To bring this project to completion, a system that will satisfy the project requirements as laid out in Section 2.1.1 needs to be conceived. Two methods are described in this thesis. First, a dual convolution method was considered; however, it was quickly found to be inadequate. Next, an Ambisonics-based solution is built, and a user experiment is designed to put it to the test and evaluate its performance.

This chapter lays out the techniques that were employed to build the experimental setup and the reasoning behind the choices that were made. Next, the user experiment is described, including its procedure and the data that we are looking to collect. Finally, the data analysis procedure is explained.

Contribution of Authors

This project was carried out individually and all the research was conducted independently. However, the author received advice from his supervisor, CIRMMT staff, and other fellow lab members, all of whom are recognized in the acknowledgments section of this thesis. In particular, Cyan Kuo helped devise the protocol for the user study and the writing for the Research Ethics Board (REB) application to obtain approval for conducting research involving humans. Prof. Cooperstock and Max Henry proof-read, commented on, and helped to improve this chapter.

3.1 Early System Design

The initial approach to generating spatial audio from a microphone signal was to convolve the stream of audio coming from the other participant with a HRTF that corresponded to the relative direction the sound was coming from with respect to the receiver's head orientation. This made intuitive sense since sound mostly comes from the direct path between the source and the listener [35]. The system would receive head tracking data, and given the orientation angle, select the appropriate HRTF from a dataset and convolve it with the incoming audio stream, giving the listener the impression of a sound source locked in space.

In order to incorporate room acoustics, we would also convolve with a RIR. This would color the audio in such a way that it would sound like it was created within that room [36]. Whether to compute the RIR or HRTF convolution first remained undecided and would have been chosen depending on preference or newfound literature.

3.1.1 Convolution Engine

The convolution between the impulse response of the filter and the incoming audio needs to be done as fast as possible, especially considering the real-time constraint of this project. For this reason, the convolution of two signals will be done through the multiplication of the signals in their frequency domain, as demonstrated in Equations 3.1–3.4.

$$y[n] = (x * h)[n] \tag{3.1}$$

$$x[n] \overset{\mathcal{F}}{\longleftrightarrow} X[\omega] \tag{3.2}$$

$$Y[\omega] = X[\omega]H[\omega] \tag{3.3}$$

$$H[\omega] = \frac{Y[\omega]}{X[\omega]}, \qquad \omega \in [-\pi, \pi] \tag{3.4}$$

Since its introduction in 1965, the fast Fourier transform (FFT) has become one of the most important tools in signal processing [37]. This algorithm efficiently computes the discrete Fourier transform (DFT), which can be used to compute the frequency representation of a discrete-time signal [38]. Today, *fast* convolution techniques make use of the FFT. Therefore, in order to compute the convolution as fast as possible to meet the timing requirements, FFT algorithms will be used.

On general-purpose processors, real-time audio processing can only be done in blocks. That is, the audio stream is partitioned into blocks or frames of audio samples, on which we can perform signal processing operations. Typical lengths for blocks are powers of 2, generally between 64 and 1024 samples. At 44.1 kHz, the standard sampling rate for compact discs (CDs), these block lengths correspond to a duration between 1.45 ms and 23.22 ms, respectively. The convolution can only start once an entire block has been received. Therefore, the signal processing cannot start until the duration has elapsed. No matter the speed of the hardware used and the efficiency of the algorithm, the latency cannot be lower than the duration corresponding to the block length. Once a block is processed, it will be requested for playback. The interval between the successful acquisition of a block and its request for playback can be utilized for signal processing. This processing typically occurs during the acquisition of the subsequent block, meaning that it must be completed within the duration of a block to avoid playback dropouts. Consequently, smaller block lengths are preferred to reduce latency; however, this approach increases the computational demands on the processor. Auxiliary operations, such as loading the block into the central processing unit (CPU), consume part of its cycle time. A CPU cycle is the basic unit of time in which the CPU performs a single operation. Consequently, only about 90% of each CPU cycle can be effectively used for signal processing [39].

Convolution is carried out by element-wise multiplication of the input with the filter in the frequency domain. Considering a block length B of 1024 samples and a filter length F, the K-point DFT will be of length $K = B + F - 1$. However, playback requires the same number of samples as the input, B. These extra $F-1$ samples are still important pieces of the audio, and simply discarding them would cause playback artefacts. Since these are meant for the next block and possibly other ones, we need to save them and apply them in the next loop.

Two algorithms, overlap-add (OLA) and overlap-save (OLS), are meant to help with this procedure [39]. OLA and OLS are preferred for short filters, when $F \leq B$. For longer filters, other methods are preferred. Using the LISTEN HRIR database, the HRIR are 512 samples in length [40]. This database contains measurements from multiple participants, giving more options for a HRTF if the spatial audio effect does not work on a user in our study. As discussed in Section 2.2.4, *average* HRTFs do not work well for many people.

Overlap-Add Method

The OLA method is illustrated in Figure 3.1. Both the filter and the input block of audio are zero-padded to a length of K. Following the computation of the FFT for both of these, they can be multiplied element-wise, which computes a convolution. By taking the inverse fast Fourier transform (IFFT) of the result, it is transformed back into the time domain. Since the result is of length K, it overlaps that of the previous convolution. The leftmost B samples can be sent for playback.

This method was used in order to compute the convolution between the HRTFs and the incoming audio. It was selected because it enables reuse of the FFT with the zero padding instead of repeated samples, which is useful when doing block computation.

Overlap-Save Method

The OLS method is illustrated in Figure 3.2. Just like with OLA, the filter is zero-padded at its end to be of size K. At each audio cycle, each block of length B is added to the end of a sliding window of size K after all the other content of the window is shifted left by B samples. The left-most B samples are discarded. Next, a K-point FFT of the zero-padded

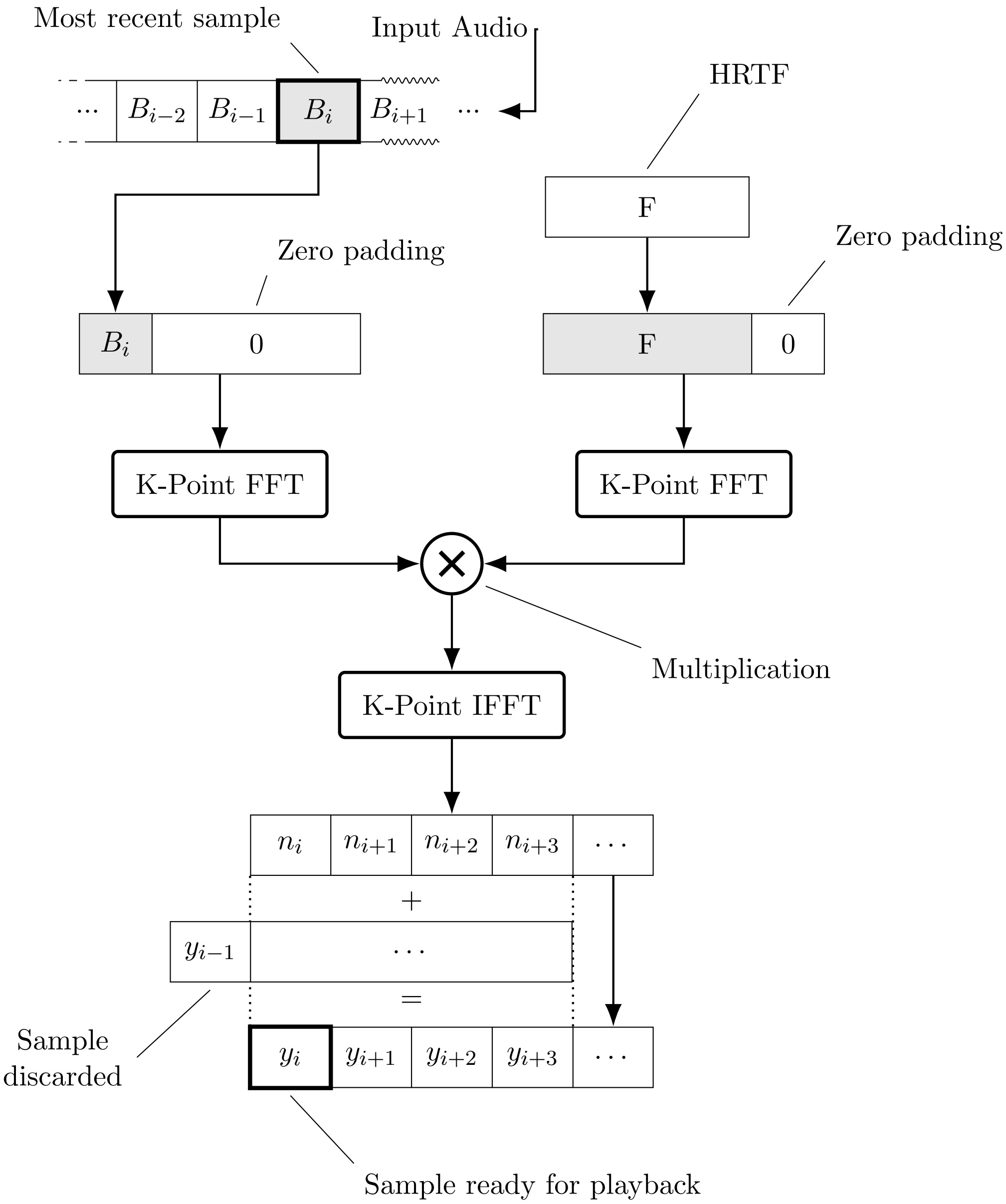

Figure 3.1 A block diagram of the OLA algorithm. Samples are convolved with the HRTF and the result from the IFFT is added to the result from the previous loop with an offset of 1 sample. The leftmost sample is then ready for playback. Figure adapted from Wefers [39].

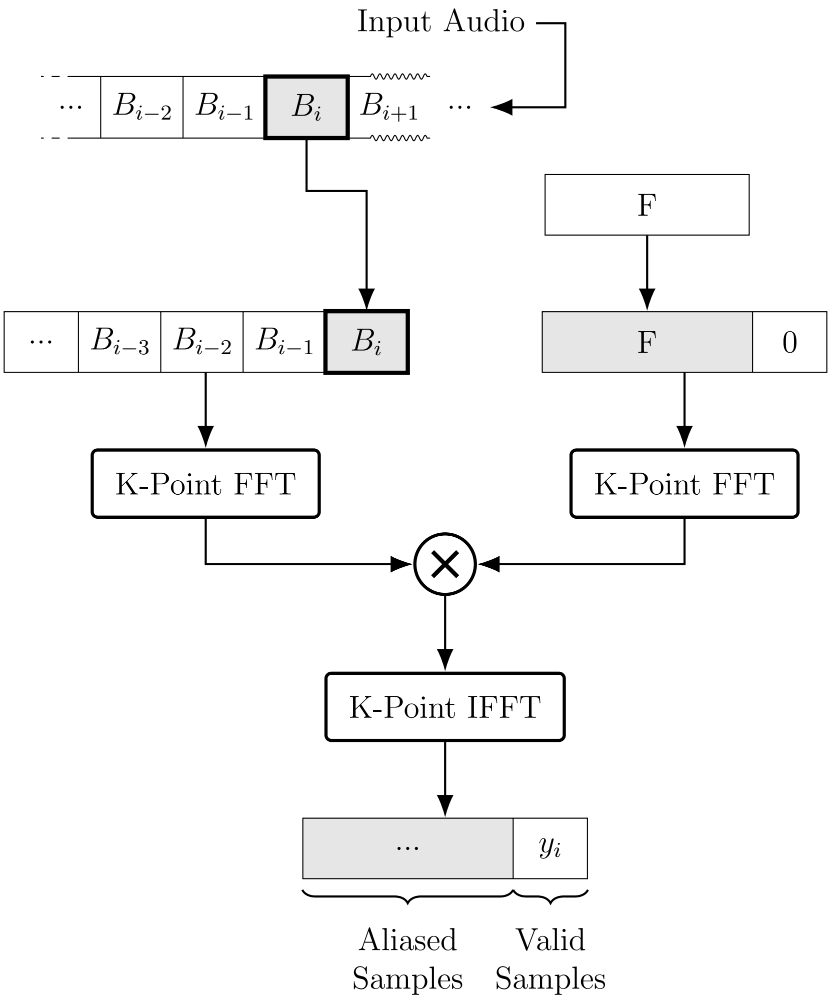

Figure 3.2 A block diagram of the OLS algorithm. Samples are convolved with the RIR and the valid samples are extracted from the result of the IFFT. Figure adapted from Wefers [39].

filter and of the sliding window is performed. Since all input values for both the filter and the audio input are real-valued, the resulting DFT spectrum is Hermitian symmetric [41]:

$$X[\omega] = X^*[(2\pi K) - \omega] \tag{3.5}$$

As a result, the DFT spectrum can be reconstructed from N unique coefficients:

$$N = \left\lceil \frac{K+1}{2} \right\rceil \tag{3.6}$$

These include the real-valued terms at $k = 0$ and, when K is even, at $k = K/2$, along with $\lfloor (K-1)/2 \rfloor$ complex-valued coefficients. From there, all N coefficients in the H (filter) and X (input) spectra are pairwise multiplied with complex multiplication. Finally, a K-point, complex-to-real IFFT is computed, and only the rightmost B samples are saved for playback.

This method is slightly more efficient than OLA since the addition is not needed at the end of the operation, so it was chosen for the convolution between the audio and the RIR. The RIRs are typically much longer than the HRTFs since reverb can last several seconds. The placeholder RIRs used during development were between 32 000 and 400 000 samples long, the former from an office and the latter from an auditorium. These lengths are consistent with expectations, since the RT60 (Reverberation Time 60) of an average residential room is typically less than 1 s and gets higher for larger spaces. This implies an RIR length of under 48 000 samples at a 48 kHz sampling rate. RT60 is the duration it takes for a sound to decay by 60 dB after its source has stopped emitting.

3.1.2 Inadequacies

Unfortunately, this design has multiple flaws, which are discussed in this section, and consequently forced the reconsideration of the methodology. Many changes were made and the final design will be discussed in Section 3.2.

Sound Reflections

The original approach overlooked the significance of reflections within a room. The HRTF tended to localize all sounds to a single point in space, whereas in reality, the initial sound wave travels directly to the listener, while subsequent reflections arrive from various directions. These early reflections typically occur in clusters, forming orders of reflections: a single one constitutes the first order, two reflections the second order, and so forth. These reflections originate from different directions and serve as crucial cues for the listener, providing insights into the spatial dimensions of the environment [42].

Late reverberation, which arrive last, contain less energy and offer fewer directional cues, appearing to emanate from all directions. Nevertheless, they are vital in conveying the room's size to the listener. Indeed, the most critical cue is the *reverberation time*. It is essential for first-order reflections to surround the listener rather than emanate solely from the source, as this spatial information is integral to the listener's perception. Therefore, the early reflections can be modeled using the image-source model as discussed in Section 2.2.4, while the late reverberation is simply an omnidirectional sound.

Singularity of a RIR

The importance of the variation in direction for the sound reflections also highlights another flaw in the original design: the use of the RIR. A RIR is inherently single point to single point, meaning that it is only accurate for a source in the location of the emitter and listener in the location of the microphone when the RIR was recorded. These also record the room reflections; however, these will be inaccurate if the locations are not matched, making this solution not versatile or flexible. Considering that this project has the ultimate ambition of enabling musicians to move within their space, this method proved to be inadequate.

This is reinforced by the fact that the process of recording a RIR is difficult. Not only is special hardware required, such as high quality speakers and microphones, but the devices they are connected to need to preferably be able to record and playback in as high of a bit rate as possible. A sine sweep is then played through the speakers and the microphone picks it up. Transient methods also exist, through the recording of a loud but short sound such as a balloon pop, but these methods are much less reliable and often produce unsatisfactory results [43]. This solution presents a barrier to widespread adoption by musicians, as the required procedure may not be accessible to those with limited resources.

The Double Convolution

The method described above relied on the computation of two convolutions in series. Although it is a linear operator, the HRTF and RIR are time-varying systems with respect to the location and orientation of the listener's head. Since both filters evolve in a correlated manner when the head moves, these operations cannot be computed separately and sequentially. If the HRTF is applied first, then we are simulating the way the input sound would be perceived by the listener with respect to their relative locations. Next, the RIR is applied, which is supposed to apply the room acoustics. However, they are being applied to this sound that is already transformed into what the ears of the listener should perceive. This means that some of the spatial cues might be removed, and it also compresses all the reverberation that should surround the listener into a single location that is encoded by the HRTF. This system behaves as if what the ear perceives was the signal being emitted into the room, as seen in the top row of Figure 3.3. If the RIR is applied first, then we are adding all of the room acoustics into the signal, including the reverberation. Next, when we apply the HRTF, we are applying the same directionality to all of the signal, so the listener will again be lacking the reverberation that surrounds them. This can be observed in the bottom row of Figure 3.3. We know that we do not perceive the environment in this manner, as discussed previously.

Overall, this means that we need a solution that applies a different HRTF to each sound wave depending on the direction it comes from with respect to the listener's head. As previously discussed, the limitation of RIRs is that they typically do not include or encode

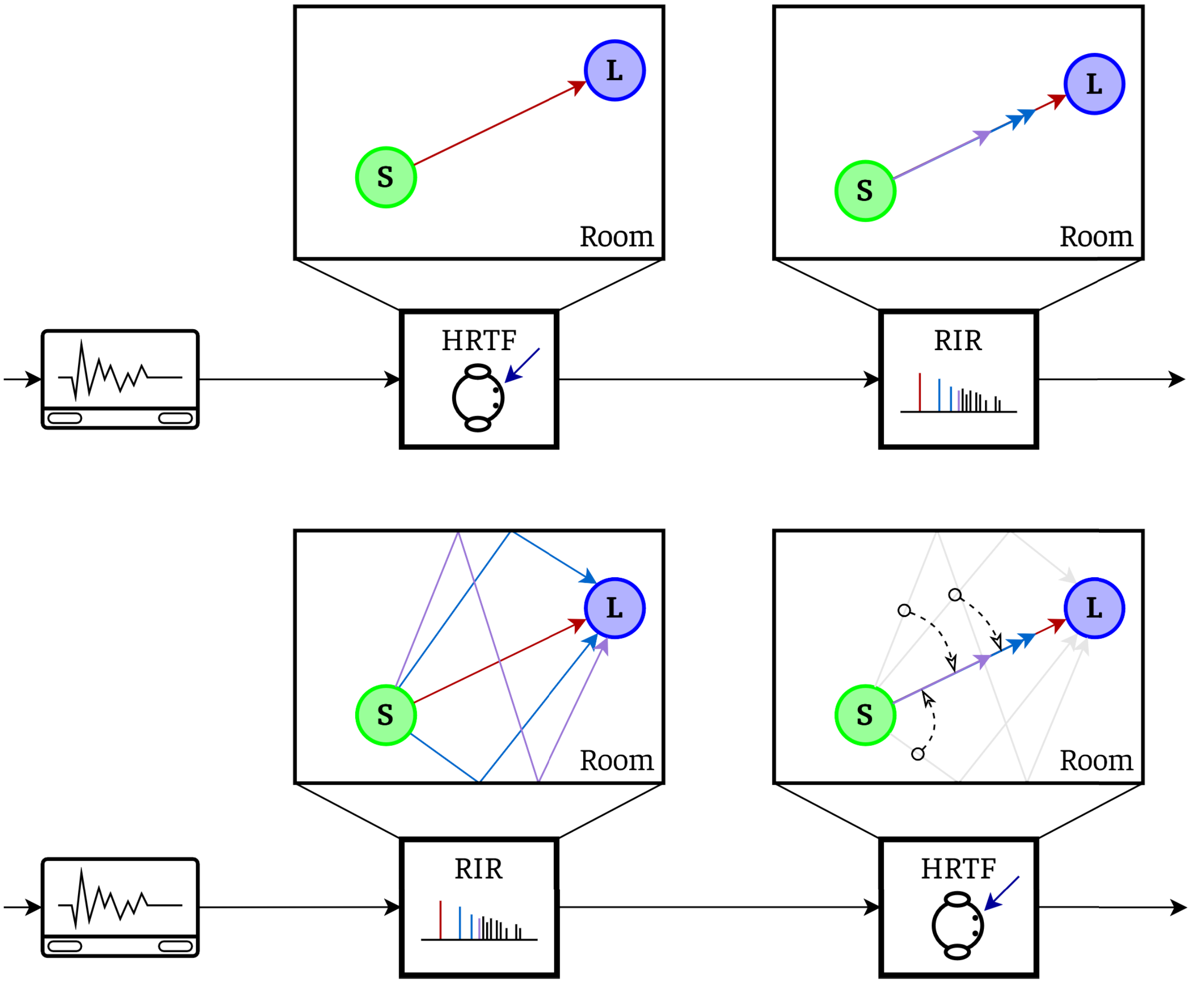

Figure 3.3 Illustration of the two possibilities for a double convolution. The top row illustrates the HRTF being performed first followed by the RIR, and the bottom row shows the reverse.

spatial information. Consequently, it is challenging to determine which HRTF should be applied to each reflection. Since they are also difficult to measure, it makes sense that we should try to find another solution. One solution would involve room modeling from which sound waves could be simulated in a manner similar to ray tracing for graphics. This method is called the image-source model [44]. It also turns out that a lot of rooms are in the shape of a parallelepiped, that is rectangular with right angles between each adjacent face. Using this knowledge, a simple but versatile model can be used: the shoe box model [45]. The new implementation is discussed in the next section.

3.2 System

3.2.1 Sound Processing

The requirements for this application were decided based on what would be appropriate for musicians. The performance of the system needs to be adequate for live performance. The aim was to keep the latency below 20 ms to 30 ms as that is the maximum delay that musicians accept for music performance [10].

3.2.2 Ambisonics

When trying to find a new solution for the audio processing, it became evident that using Ambisonics would be the answer. Ambisonics is a surround sound format that works by modeling a sound field instead of individual speaker channels [46]. It is also a specific recording technique that requires special microphones to capture the sound field.

The reproduction of the sound field can be done with an arrangement of loudspeakers, typically less than what is required in *traditional* surround sound formats [47]. The sound field, denoted $S(\mathbf{x}, \omega)$, can be synthesized by the loudspeaker arrangement according to the following description

$$S(\mathbf{x}, \omega) = \sum_{l=0}^{L-1} \hat{D}(\mathbf{x}_l, \omega) \cdot G(\mathbf{x} - \mathbf{x}_l, \omega) \tag{3.7}$$

where L corresponds to the number of loudspeakers, $G(\mathbf{x} - \mathbf{x}_l, \omega)$ the spatial transfer function of the loudspeaker at position $\mathbf{x}_l$ and $\hat{D}(\mathbf{x}_l, \omega)$ its driving signal. The sound field is being controlled at the origin of the coordinate system. The multiplication of $\hat{D}(\cdot)$ with $G(\cdot)$ results in the sound field that is emitted by a given loudspeaker. The addition of all sound fields, or superposition, produces the overall sound field $S(\cdot)$. Therefore, we can choose the appropriate loudspeaker driving signals $\hat{D}$ to obtain the desired sound field $S(\cdot)$ at a given point. Each of these quantities are expanded into surface spherical harmonics whose coefficients depend on the order. The higher the order, the more components (called modes) in the decomposition of the wave propagation. Orders of two or higher are called higher order Ambisonics (HOA) and possess modes for plane wave propagation towards the origin. A higher order enables higher resolution and makes a larger sweet spot possible.

We therefore have a system of linear equations that are referred to as *decoding equations*. This makes the format very flexible as one can just decode the sound field as appropriate for their speaker setup. This is a process called wave field synthesis, which involves solving the equation to reproduce the sound field with the available speakers. In our case, we can use Ambisonics to apply room acoustics since, using a model that will be described next, we can encode sound as coming from directions around the listener based on their position and that of the source. Furthermore, multiple plugins for our digital audio workstation (DAW) of choice, Reaper [48], are available to work directly in this format [49]. From there, it is easy to decode into binaural audio for the headphones that the participants will be wearing.

3.2.3 Room Model

A sound that was produced in a room will be affected by its environment. Humans are good at creating a mental model of the way something should sound given a visible environment. Therefore, to give the impression that a sound was created within the environment that the musician is in, a model for their room needs to be created. These models can be extremely complex, where every surface can be accounted for with their own specific absorption spectra and diffusion pattern. However, such a detailed model is not necessary to create the illusion of a sound source within the environment. In fact, extreme approximations can be taken with surprisingly good results. According to the ISO-3382-1 (2009) standard, the parameters to characterize room acoustics are reverberation time (RT), Early decay time (EDT), clarity indices C80 and D50, sound strength (G), and the interaural cross-correlation coefficient [50]. The JND for RT across participants can vary by over 30% [51]. As for clarity, one study found the C80 JND to be 4.4 dB [52]. Moreover for the EDT, the JND was found to be 18% [53]. This demonstrates that for a participant to be unable to distinguish between the acoustic characteristics of the model and those of the real room, the parameters do not need to be matched with perfect accuracy to the actual room parameters.

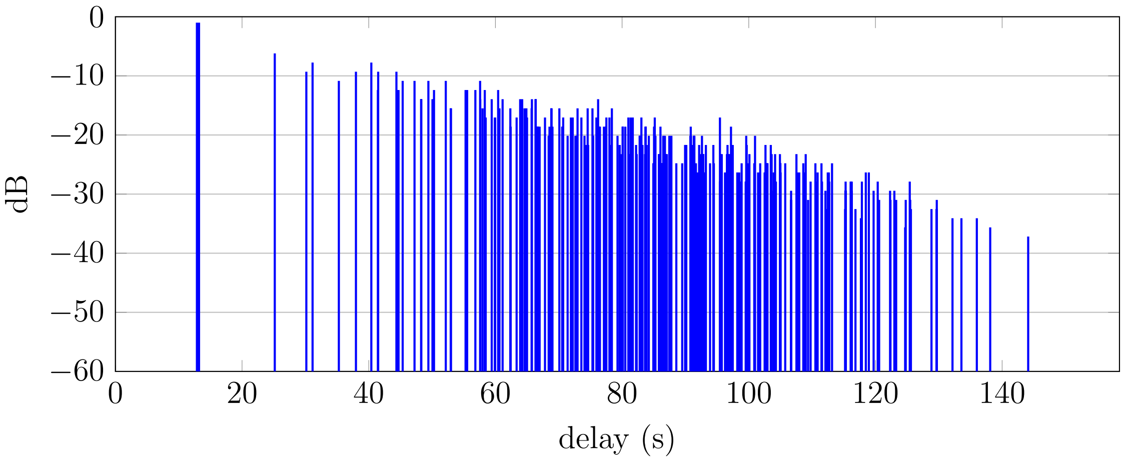

Figure 3.4 A sample RIR. The direct path can be seen as the first impulse, followed by the early reflections, then the late reverb. This simulation was limited to around 300 reflections.

A common approximation is the shoebox model. This model is for rectangular parallelepiped rooms, that is, a rectangular room with all adjacent surfaces orthogonal to one another. Given a source and receiver location within the room, it is easy to compute the paths of sound waves up to a certain order, meaning a number of reflections, using the image-source model as described in Section 2.2.4. The direction, delay, and attenuation of the signal can be obtained at the listener location. Since we are using Ambisonics, we can obtain the sound field where the listener is located. Figure 3.4 shows the RIR of a simulated room with a source-receiver pair. We can observe the first impulse representing the direct path from the source to the receiver, followed by the early reflections. These are all followed by the dense late reverb, which is not shown in this figure as it is not modeled by the shoebox model. Here, the reflection simulation was limited to 256 reflections, so the tail of the RIR is not complete. However, using a simple reverb plugin, we can add the late reverb to the room simulation. The late reverb has reflections so dense that it can be considered omnidirectional [54].

To achieve this, the DAW Reaper is used [48]. Many plugins for Reaper are available, including the Sparta suite from Aalto University in Finland [49]. This suite contains many tools for working with Ambisonics including ambiRoomSim, which is used in this system to simulate the room for the musicians. It supports eight third-order receivers, meaning that the plugin can support up to eight musicians physically in the same room using a third-order Ambisonic signal. This is useful since ultimately, this project aims to simplify shared experiences, and this enables not only single musicians in separate rooms but also multiple musicians in multiple rooms, creating a mixed-reality environment. This plugin also supports the displacement of sources and receivers, to some extent, enabling musicians to physically move around their space, and the sound simulation will change accordingly. For now, however, the tracking is only of the head orientation, as described in Section 3.2.4, so this feature is not evaluated here.

The virtual source in the ambiRoomSim plugin was placed in the location of a speaker in the room to make use of the ventriloquist effect, as described in Section 2.2.5. The aim is to help musicians anchor the sounds they hear to their physical location within the space. Given that they will hear other musicians without visual cues, they may find it challenging to localize the sound within the room. To overcome this, a speaker will be used as a proxy source, creating the illusion that the sound is emanating from a specific location. This approach is expected to facilitate the musicians' ability to anchor the perceived sound sources within the space, thereby improving their spatial perception of the audio.

The location of the receiver in the ambiRoomSim plugin matched that of the participant in the physical room.

3.2.4 Head Tracking

For the sound to be processed according to the position of the user, their head needs to be tracked. This means understanding the location of the head within the room, as well as the direction the head is facing. There are several methods available for tracking head movements, one of which is commonly known as motion capture or mocap. Mocap can utilize active markers, as in electromagnetic motion capture, or passive markers, as in optical motion capture. The latter method requires the use of cameras. However, most of these systems are very expensive or bulky, while the aim is to make ours as accessible as possible. One motion capture technique uses inertial measurement unit (IMU) sensors, which measure linear acceleration and rotational rate. Obtaining absolute position and rotation from acceleration measurements is subject to double integration, and therefore amplifies drift [55]. Some more advanced IMUs, sometimes known as 9D IMUs, also measure the magnetic field and combine the data from all sensors to try to mitigate these errors in what is known as sensor fusion [56]. Kalman filters have often been used for their high estimation accuracy and robustness to poor measurements due to high activity or fluctuating local magnetic field. However, one recent and more capable algorithm is known as the Madgwick algorithm [57]. This reduces the drift when tracking using an IMU device, similar to a traditional motion capture method, but IMUs are cheap and very small.

To make the system accessible to most people, the original experiment design consisted of strapping a cell phone to the head of the user. Phones nowadays consistently have IMUs on board and generally implement a good sensor fusion algorithm as well [58, 59]. From there, it is only a matter of transmitting the sensor data from the phone to the computer, which can be done wirelessly over Wi-Fi using the Open Sound Control (OSC) protocol. This makes the system easily accessible to most people since the majority of people, particularly in developed countries, have smartphones [60]. However, during pilot testing, it was apparent that the phones, two iPhone 12 Mini devices, were too heavy and therefore uncomfortable for the musicians.

As a result, a new device was designed to replace the phones. Considering the ultimate goal of the project is to implement some sort of headset for mixed reality and that these generally have IMUs built into them [61], the device that is built for this experiment did not need to be appropriate for long-term deployment, only sufficient for the purpose of head tracking in a repeatable study. Until such a solution is implemented, a smart watch could also serve as the head tracker due to its lower weight compared to the phone and relative market penetration, though less so than phones. Such a device was not tested since it will eventually be replaced and, therefore, it made more sense to use a lightweight, low-cost, and readily accessible device to carry out the head tracking. A Raspberry Pi Zero W device running Raspberry Pi OS was selected for cost and ease of deployment since these can connect to Wi-Fi and also use GPIO. For the IMU, an Adafruit BNO055 device was selected since it performs sensor fusion on device and can communicate over I2C with the Raspberry Pi. The BNO055 is soldered to a prototype board that was mounted below the Raspberry Pi with standoffs. A third layer of prototype board was mounted just below in order to provide a flat surface to tape to the top of the headphones. A power plug was mounted to the side to provide power to the device, with a wire running down the left side of the head, bundled with the headphones cable, down to a USB power adapter, providing the device with 5V

DC. While wire powered at present, it could be battery powered for a little additional weight but increased freedom of movement.

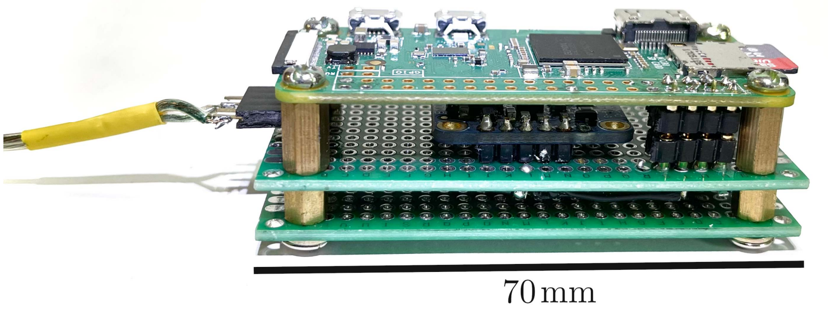

Figure 3.5 The Raspberry Pi-based head tracking device.

The resulting weight of the device is only 41 g, compared to 135 g for the phone. Further optimization could have been achieved, as the brass standoffs contribute significantly to the overall weight. The device can be observed in Figure 3.5. The author programmed the Raspberry Pi to retrieve the IP address of the main computer upon startup and automatically start broadcasting yaw, pitch, roll data to it. The host computer receives the OSC data from each participant and a Python script redirects the message to the corresponding track effect in Reaper. This script also supports aligning the orientation from the IMUs to that of the room. Since the devices start automatically, their reference frame, defined as *roll*, *pitch*, and *yaw* all set to 0^o, can be misaligned with the reference frame of the room. The script, when triggered by a special OSC message sent from a phone, can record the current orientation and subtract it as an offset from subsequent readings, aligning the data to the room. The device is oriented such that it faces one end of the room, and the script can save the direction as the offset so that it is aligned with the virtual room in Reaper. This is done by sending a special OSC message to the script that saves the current orientation of the IMU and uses that value as an offset.

3.3 User Study

In order to validate the design of the audio experience for the Music Telepresence project, a study was designed to understand what aspects of the experience users appreciated and possibly benefited from, as well as the factors they found unappealing or problematic. The study was approved by McGill University's Research Ethics Board, REB #24-02-115.

3.3.1 Experiment Design and Setup

To evaluate the performance of the system and obtain feedback from musicians, a study was designed. Two participants will play music together while in separate rooms. No restriction was placed on participants other than the need to have normal hearing capabilities. Their instruments are picked up using an appropriate microphone set-up. In the case of pianists, a keyboard was used and wired into a loudspeaker placed behind the keyboard and into the audio interface so that the pianist had their instrument resonating in their physical room and also sent out to their partner. Each participants wears open-back headphones in order to hear their partner while also being able to hear their own environment, in this case the Senheiser HD-595. Each mono microphone is wired to an RME Fireface UC audio interface. The sound processing happens in Reaper as described in section 3.2.1. This produces a stereo sound output sent through the audio interface to each musician's headphones, hence the double arrow in Figure 3.6, where the one-way audio path can be seen in a block diagram.

Participants were given 15 minutes to get accustomed to the music piece and setup. After the warm-up period, two conditions were tested: flat, which is the control, and binaural.

The flat condition uses no processing. The sound that is captured by each microphone is directly sent to the other participant's headphones. This is the condition that everyone is used to, since usually, when listening to music with headphones, no spatial effect applied to the sound, which is directly played back in the listener's ears. The same applies to musicians on the occasions that they wear headphones, often for monitoring purposes. Therefore, this condition acts as a control since it will sound the way that participants will be most used

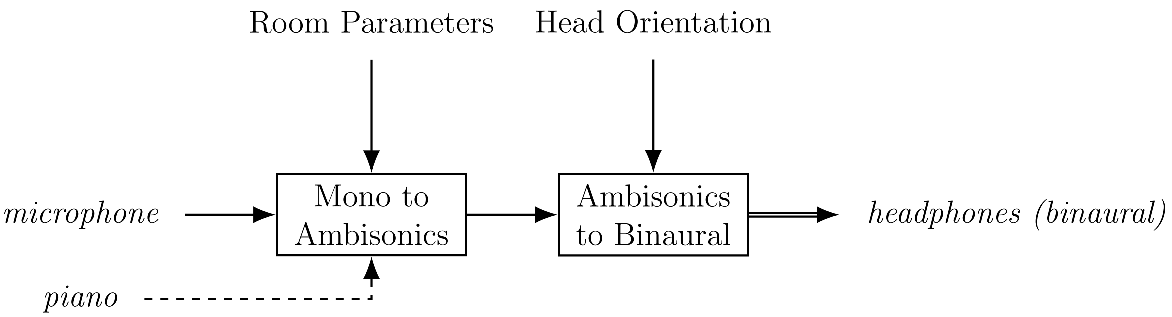

Figure 3.6 Block diagram of the audio path for the user study. The microphone and the piano (for participants that used it) mono signals are transformed into Ambisonics according to the room parameters of the receiver. Then, the audio is transformed into a binaural stereo signal according to the head orientation parameters of the receiver.

to.

In the binaural condition, the sounds produced by musician A is processed as described in Section 3.2.1 in a manner that sounds like it was produced in the room where musician B is, and vice-versa. The headphones have an IMU attached to the top of the band as described in 3.2.4 that informs the sound processing, creating a natural-like experience where the sound source is fixed in space and is perceived in a way that is characteristic of the physical space the listener is in.

Participants were offered a CAD15.00 per hour compensation for their time. Experiments were estimated to last between 1 and 2 hours. This estimation proved to be correct as most sessions lasted about an hour and a half.

3.3.2 Procedure

Subjects were solicited through a CIRMMT mailing list, a poster located in McGill's Elizabeth Wirth Music Building by the elevators, and a post in a Facebook group for people interested in participating in paid studies. During recruitment, participants that showed interest in participating in the study were sent an introductory questionnaire in which we asked about their music-playing background. Based on their answers, participants were selected if

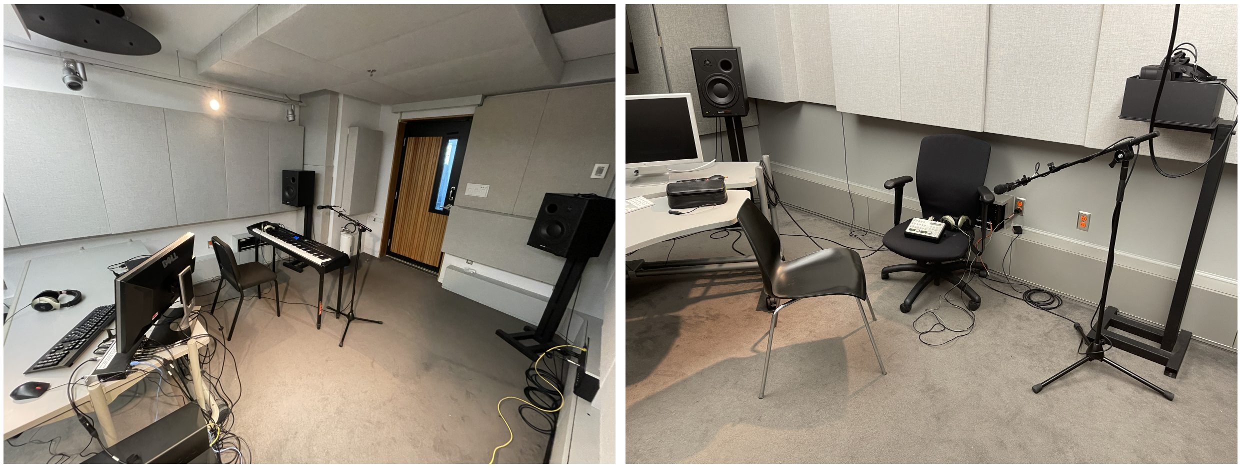

Figure 3.7 Room 1 (left) and room 2 (right) for the experiment. In room 1, the keyboard was moved out of the way if it was not being used and participants typically faced 90° to the right of the piano position. The speaker seen on the right was in the location of the virtual sound source from which the other participant was heard. The computer running the audio simulation is also in this room. In room 2 were the participants that did not play the piano if one was present. The speaker seen on the left was in the virtual source location.

they stated to have been playing an instrument for a minimum of 7 years. This is to ensure sufficient familiarity with their instrument and high music playing abilities, in order to make sure the difficulties the musicians were having were because of the setup and not because of their struggles to read the music sheets or to play the song. Once selected, participants were paired up, either by acquaintance or by instrument compatibility. This usually meant avoiding two piano players playing together. Once a match was made, scheduling attempts were made and they were then invited to the study space at CIRMMT in Montreal.

Participants were greeeted in the lobby and brought to the study rooms. They were shown around the space, specifically the chair for them to use, the microphone they would be heard through, the headphones they should use, and the volume knob if they wanted to adjust the loudness of what they were hearing. They were also explained how the experiment would go and they were given the music sheet for the piece they would be playing. They were then invited to get comfortable and set up to play their instrument. Prior to starting the experiment, each participant was interviewed to gather demographic information and insights into their familiarity with spatial audio. After setting-up, participants were given a 15 min-period to warm up and get used to playing together using the system. Some participants knew each other and played a piece they were both familiar with, while others were given an adaptation of Ode to Joy by Beethoven for their instrument. This piece was chosen for its ease of playing and its fame since we are evaluating the collaboration of musicians, not their individual playing abilities.

Once both participants were ready, they were asked to play for about five minutes in run. Three runs were carried out, with at least one of each condition, and the order was balanced across participants. After each condition, the participants were asked a few simple questions to reflect on their experience. Finally, after the last condition, they were asked questions pertaining to the experience as a whole in order to evaluate their perception of the differences between the conditions, their preferences, and how it compared to their usual music-playing habits. The interviews are semistructured, with the guiding questions found in Appendix A. Once the last interview was completed, they were invited to gather their belongings, they were paid, thanked for their time, and walked to the door of the research institute.

3.3.3 Participants

Following the procedure outlined in Section 3.3.2, 5 experiments were carried out with 10 different participants. A selection of relevant demographic data including the instruments that the musicians played are listed in Table 3.1.

3.3.4 Data Analysis

A reflexive thematic analysis was performed on the data that was collected during the experiment, following the method laid out by V. Braun [62].

Demographics								
Gender	4F	6M						
Age	Min	Max	Mean					
	20	40	28.1					
Mean years of experience	16.1							
Instruments	Bass	Cello	Clarinet	Guitar	Piano	Saxophone	Viola	Violin
	1	1	1	1	3	1	1	1

Table 3.1 Participant demographics for the experiment.

Transcription

All interviews were audio recorded and notes were taken during the experiments and the interviews. From the audio recordings, transcripts were created using an on-device speech recognition tool called WhisperX [63]. The model "large-v2" was used since it is recommended by the author of the tool. This tool also performs speaker diarization (the process of identifying and separating different speakers in an audio recording), which is useful for analysis as it organizes the transcript neatly. A new line is written on whenever the speaker changes, labeled as speaker 1 and 2, and from the context, it is trivial to replace these by *interviewer* and *participant* afterwards. Once the transcripts were created, they were cleaned up since the diarization was sometimes imperfect and a few sentences could use some additional punctuation to reflect the speaker's intentions such as hesitation or pauses. The tool also does not pick up "umm" and "uhh" and so these were added back in manually.

Coding and Tagging

With a text file produced for each interview, each was loaded into a local instance of Taguette [64], a qualitative research tool for tagging text files and keeping track of tags. Following Braun's method for thematic analysis, the first step was to get familiar with the data. Although the author had participated in each experiment, reading through every transcript and giving them all an equal amount of attention allowed for a comparable level of recollection for every conversation. This also enabled the search for patterns and similarities across some or all participants while keeping in mind the subject of the research, whether spatial audio with room acoustics had a positive impact on the musicians, and what aspects of the experimental setup need improvement. The initial similarities were written down for further analysis and as potential initial codes.

Common topics such as the desire to see their partner or the issues with the latency were brought up and were therefore assigned codes. Certain codes were identified as overlapping, including references to the challenges of communicating with partners and discussions about collaborating and synchronizing with them. This overlap resulted in both positive and negative aspects of communication with the other musician. The initial codes that were extracted were communication, setup issues, sound, spatial audio, timing, visual/line of sight, remote play, and interesting as the catch-all for anything else that could be relevant but not classified yet.

Chapter 4

User Study Results and Discussion

In this chapter, we discuss the three main themes revealed through our reflexive thematic analysis, which will help answer our research question. First, the ‘audio experience’ theme recounts the evaluation of the sound, spatial audio effect, and latency by the participants. Next, the ‘remote collaboration’ theme describes the experience of the participants when it comes to playing remotely. Finally, through the ‘technical and setup shortcomings’ theme, we discuss the ways that the experimental setup did not address some of the needs or hindered the playing of the participants. These themes give promising insights regarding the use of room acoustics and spatial audio in remote music performance. By examining these findings in the context of existing literature and considering their practical applications, we aim to provide a comprehensive understanding of the benefits and challenges associated with the use of spatial audio in remote music performance applications.

Contribution of Authors

The data analysis methodology was selected with the help of Juliette Regimbal, who also advised on the Bias section. Prof. Cooperstock and Max Henry proof-read, commented on, and helped to improve this chapter. The analysis and the discussion in this chapter is the work of the author, guided by the methodology laid out by Braun [62].

4.1 Bias

Prior to discussing the results obtained in the previous chapter, it is important to acknowledge the potential for bias in the analysis that was performed as it is qualitative [65, 66].

My background in music and familiarity with audio technology may have influenced my interpretation of participants' experiences with spatial audio and remote collaboration. I received six years of music education in France, focusing on both instrumental performance and music theory. However, I have not actively engaged in musical practice over the past decade and do not identify as a musician. Additionally, my scientific mindset and 6 years of engineering studies may have shaped the experimental design and the questions posed to participants.

As the sole researcher, my solution-oriented approach could introduce confirmation bias. My presence during the study might have influenced the behavior of participants, potentially leading to different actions than in a natural setting. The experimental context itself may also have affected their behavior.

4.2 Audio Experience

The theme of 'audio experience' encompasses the perceptions and evaluations of the spatial audio effect as mentioned by the participants. It is central to understanding how spatial audio influences the quality and immersion of remote musical collaboration. This theme reveals that while spatial audio is appreciated for its immersive qualities and participants benefited from it, technical issues such as timing, synchronization, and sound delays remain significant challenges for a distributed music performance.

4.2.1 Benefits of the Spatial Audio Effect

Participants generally expressed that the spatial audio effect was beneficial to the auditory experience, noting that it enhanced their ability to localize sounds—and therefore their partner—and created a more immersive environment. This is evident in comments such as:

> P2: The sound felt more clear, like they were playing in front of me, or like to the side of me, like [. . .] they were less in internal and more like right next to me. Which I liked.

and

> P5: I think the audio quality and the spatial audio, and I feel like he's, it feels like he's in the room. Like it's clear. I can hear what he's playing and I can hear his voice quite well. And it's quite, it's almost trippy. It is quite realistic.

These responses indicate that spatial audio can improve the listening experience by providing a more natural and engaging soundscape. Participants found that the spatial distribution of sound helped them better place their partner in the space they were in and felt more connected as a result. When asked which condition they preferred, 60% of participant indicated a preference for the spatial audio condition.

Participants also recognized the benefits of spatial audio in enhancing the remote collaboration experience. They felt that the more immersive environment could improve the overall quality of the collaboration:

> P5: (comparing the control condition to the previous spatial audio condition) it felt less real, like spatially, and audio wise real as if you were in the room. [...] it felt more like playing with a recording

> P5: I find it much more connected and much more in the music [when] playing with people live and having audio coming from different spaces than for example, to a recording. And obviously there's more energy and stuff going on there, but still, I think that spatial audio is part of that.

> P10: It was more three-dimensional and the directional quality was much stronger in that in that version. So it felt more physical, more tactile. It really felt like

> [partner] was there in the room. [...] I found that our phrasing was, we phrased together and played more dynamically. [...] It's just a more engaging feeling of playing with another body in space. I feel a bit more connected to my playing.

These quotes suggest that spatial audio can enhance the playing experience and make remote collaboration feel more natural. By applying the rough sound characteristics of the room to the incoming audio stream combined with head tracking, the musicians can feel more engaged, leading to a better experience for them.

It has been shown before that spatial audio with head tracking improved immersion and connectedness in situations where users need to communicate [67, 68]. From the interviews with the participants, it is clear that they also felt that the effect was beneficial, with some mentioning that they felt more connected, they could place their partner in space, and they felt like they were in the room with them.

The spatial audio effect worked in tandem with the room acoustics simulation, so from our experiment, it is difficult to assess the individual contribution of each of these two components. However, it has also been shown that incorporating room and environment acoustics into virtual environments increased presence and engagement [69]. Therefore, our results confirm our hypothesis and what is found in the literature. In fact, it was immersive to the point that two participants reacted as if they believed the audio stream was coming from the speaker in the corner, rather than from the headphones, when the spatial audio effect with room acoustics was activated:

> P2: Oh I don't need these [headphones] for this run

> P8: Do I listen to her through the headphones or the speakers?

This is encouraging because it shows that the effect can work so well, participants truly believe that the sound comes from the room they are in, meaning the acoustic requirements for presence are fully met, at least for some participants.

However, some of the participants gave no sign of noticing the effect or a convincing description of the differences between the two acoustic conditions. This could be due to the HRTF used by the binauralizer plugin, which might not be a good fit for that particular participant. A future iteration could look at *simple* methods for personalizing HRTFs, though a state of the art method would be an individualized measurement for each ear for each participant [28].

4.2.2 Latency and Synchronization

Despite the benefits and the appreciation for the spatial audio by the participants, they also encountered challenges related to timing and synchronization. In the spatial audio condition, the system latency is 42.4 ms, an increase of 34.6 ms over the control condition. Although the impact varied among participants, the majority reported challenges with synchronization, which hindered their ability to play together as effectively as they did in the control condition.

> P2: In the second recording session, I felt like I was rushing. But to him, it sounded like I was slowing down.

> P9: both the first and the third [conditions] made me feel very strange playing music, feelings that I don't really get often because of how difficult it was to keep time, and the fact that you're slowing down really messes with your head and whether you're in the right.

These quotes highlight the importance of accurate timing and the effect that extra latency can have on the ease of playing of the musicians, even when it is small enough that they cannot quite tell that it is present. Latency in the audio can disrupt the flow of the performance and take their focus away from the music they are making by forcing them to focus on the timing.

This latency is a major drawback of our design, as mentioned by 90% of the participants. While this additional delay should have been added to the control condition in order to make them more similar, we also heard some very positive feedback about the responsiveness of the control condition. This gives great motivation to attempt to reduce the latency added by the system since making the latency similar to the control condition would enable the musicians to “lock in and have fun playing together” (P5), as they described the control condition. Participant 5 in particular played a very rhythmic song, making them more susceptible to noticing the extra latency. The latency of the control condition being 7.8 ms and the spatial audio condition 42.4 ms, they are on either side of the limit of 30 ms before noticeable slowdowns begin to occur [10]. This latency was clearly a barrier to musical enjoyment as participants mentioned having to focus more on their timing as opposed to the enjoyment of their piece.

4.3 Remote Collaboration

The theme of ‘Remote Collaboration’ explores the challenges and benefits participants experienced when playing music remotely with another musician. This theme is crucial for understanding the ability of the participants to collaborate despite the remote nature of the experiment. We are also looking to understand the effect of the spatial audio effect on the collaboration between the musicians.

4.3.1 Communication and Interaction

Effective musical communication is a cornerstone of successful collaboration, and participants highlighted the importance of being able to hear and interact with each other clearly. Musicians communicate in multiple ways, both verbal and non-verbal, and the ability—or lack thereof—to communicate was often mentioned as something important for the player to figure out:

> P2: I thought it was an interesting experience to one, play with another person in duet style in this way and figuring out how to communicate the best

> P7: [I disliked] the fact that I cannot give comments in the moment that they happen. If it passes, then it's passed. You cannot make quick comments on the things that are not right.

These quotes underscores the importance of having the ability to communicate to facilitate collaboration. While participants appreciated being able to hear each other clearly, they highlighted the need to adapt their communication channels and their difficulty in communicating quick, in-the-moment information. Musical communication between musicians is a complex interaction that will be discussed further in the next section.

Musical Communication

Musical communication refers to the cues and interactions that musicians use to coordinate, which is an essential part of musical collaboration. These cues can be verbal or non-verbal, the latter often consisting of body language, facial expressions, eye contact, musical cues, and other physical movements [70, 71].

Some of the verbal communication is instructional. This is when the musicians communicate to give each other specific instruction such as when to start playing, where to start playing from on the score ("Let's restart from bar 8"), how to play ("Try with the crescendo"), etc. This is often used in practice or rehearsal settings.

Another type of verbal communication is cooperative. As discussed by Seddon [70], this is a more democratic mode of communication than the instructional kind since musicians will often discuss together to agree on changes to make, clarify, or evaluate various problems. This type of conversation happens every time the playing is suspended through a verbal medium.

Finally, there is verbal collaboration, where musicians democratically discuss creative changes made to their performance, such as tempo or dynamics. This type of communication also takes place when the playing has stopped.

The same three types of communication also exist in non-verbal form. Non-verbal instructional communication takes place when musicians read the score with intent, which instructs them how to play, and when a musician demonstrates how to play a part to another member by either playing it or vocalizing it.

Non-verbal cooperative communication takes place while the musicians are playing and are "sympathetically attuned" and generally manifests through body language, facial expressions, eye contact, musical cues, and other body movements [70]. This mode of communication is used to help musicians stay in time and help them play together. When music performance breaks down, the playing generally stops and verbal cooperation or collaboration is used to address the issues.

Non-verbal collaborative communication is used to provide empathetic creativity and occurs through the music itself and body language. The latter is often exaggerated to express enjoyment and positive feedback regarding the collaborative performance.

Most of these types of communication have taken place during the experiment. Participants have verbally instructed each other to restart playing from a specific bar, they have also tried to smooth out difficulties they had with certain sections, and some have even discussed dynamics and tempo, despite this being for an experiment and not a staged performance. We have also observed vocalization of parts of the score, some participants commented on the embellishments played by their partner, and some were also observed gesticulating as they were playing. This last part is very telling of how they were attempting to communicate non-verbally cooperatively or collaboratively; however, their partner had no way to receive this information and to subsequently interpret it.

In our experimental setup, the focus was on the audio experience, and the results are promising. Communication through that channel appeared unimpaired, participants reported experiencing good clarity, and they were able to communicate verbally as well as musically. However, a significant portion of the communication also happens through non-auditory channels, most notably through visual means. As discussed above, body language, eye contact, and exaggerated gestures can all be used by the musicians to communicate while they play or perform.

4.3.2 Lack of Visual Feedback

Playing music remotely presents several challenges that were not addressed by our current experimental setup. This mainly included the lack of visual cues that almost every musician mentioned being a shortcoming, particularly as the visual channel is a major communication channel while playing.

> P3: there is a bit of a disconnect not seeing the person that you're playing with

> P4: I'd, like, listen-or, like, look at them to see when they're breathing and stuff, and, like, look behind me or look in front of me to see the conductor and actually be able to see them face-to-face is pretty valuable

> P6: If someone is in front of you playing, you can do a lot with visual cues and they can hear you breathing and stuff.

> P7: It's not only breathing that gives some clues of when we should start the piece. Because, for example, I do this [gesture] and he knows I'm going to play. [...] Also we cannot stop each other. [...] if I could point to the places that needs to be up [on the music sheet] because in our practice in person we always [say] you should practice more this this bar or something

> P8: In rock, it's really important to see the movement because it's kind of like the energy increases, right? Like the physical energy can impact the way you play.

These comments highlight the importance of visual feedback in music collaboration for a variety of different reasons. Without visual cues, participants found it more difficult to synchronize their playing and respond to each other's musical expressions. The absence of visual feedback can lead to a sense of disconnect, making it harder to achieve the same level of engagement as in a face-to-face setting.

Design Justifications

We anticipated the lack of visual feedback to be an issue for the participants when designing the experimental setup; however, we decided against providing a solution to this problem. This is because the Music Telepresence project as a whole has many components, and one way to divide it up to test parts individually was to evaluate the audio experience independently from the visual experience. Furthermore, the visual experience is not ready to test since we wanted to perform volumetric capture to add in the other performer in augmented reality.

When designing the study, we discussed within the Shared Reality Lab whether providing a simple alternative would make sense, such as a screen in each room displaying the live feed from a camera in the other room. We ultimately decided against it as we feared it could skew the results with participants providing feedback that would be influenced by the video feed, which could have had its own latency, low quality, or other parameters that could have influenced the perception of the participants such as the inability to recreate eye-contact. Finally, by removing the video feed, we believed the participants would be much more sensitive and pay more attention to the sound, which is ultimately what we wanted feedback on.

Effects

Through the analysis of this theme, it is clear that the participants missed the visual channel for their musical communication, as we had anticipated. Most participants made comments about their desire to see their partner, ranging from explicit and direct questions asking why we did not provide a screen, to comments hinting about struggling to communicate due to the lack of visual connection. Participants were unable to communicate enjoyment through body language or difficulties and the desire to stop playing through eye contact or other gesture, which one participant felt strongly about, citing their inability to “stop each other”. It makes sense that, as a result, participants would qualify their experience as “lacking [. . .] connection”, which the literature confirms, having high quality visual feedback improves immersion and presence [68].

It is therefore evident that the next iteration of this project should aim to provide a solution to this problem. The requirements will be challenging since the video will need to not only be low latency but also synchronized to the audio, as a participant pointed out. Failing this could lead to other issues, such as the musicians not knowing whether to follow the audio or the video.

An interesting note is that multiple participants “enjoyed the challenge” of not being able to see their partner. Note, however, that they still called it a challenge. Despite the participants all mentioning they would like to see their partner, this may be because most participants played with their partner here for the first time. It appears that once participants get used to the telepresence setup, they eventually adapt and feel like they might not need the visual feedback anymore [72]. This therefore brings the question, are we trying to emulate in-person-like music performance, or is providing a system that meets the same needs but in other ways acceptable?

4.4 Technical and Setup Shortcomings

The theme of ‘technical and setup challenges’ highlights the various technical issues and problems due to the experimental setup that participants faced during the remote collaboration experiment. Addressing these challenges is crucial for improving the overall experience and ensuring that the benefits of spatial audio can be fully realized.

4.4.1 Experimental Setup Issues

Sometimes, the experimental setup got in the way of the participants. These issues are varied, and also affect some instruments more than others. Here are some examples:

> P7: The problem is that I cannot hear myself enough. And it's not the volume, because as soon as I have the hands free, then it blocks the violin's sound a little bit. And it's a little bit more important for the violin, because, for example, this piece has a lot of staccatos, spiccatos, which I couldn't hear if I'm doing it the right way because of that.

> P7: I get used to the setting really fast, except for the wire, which goes from the left side of the ear. So if the headphone was wireless, and I could move around the room. It was better because I move around the room when we are practicing in person.

> P2: The volume was fine where I set it but I could tell that if it was a little bit higher it'd probably clip

> P1: The headphone is not really fitting well. [. . .] Maybe a better piano or a lower chair.

These quotes highlight some of the technical issues faced by the musicians. While none of them is absolutely critical, the next design iteration should take these as recommendations for improvements to try to provide as good of an experience as possible.

4.5 Areas for Improvement

One of the aims for this experiment was to gather information for what would need to be improved in a second version. After our analysis, a couple of major points have emerged as needing improvement.

Reducing Latency

As many of the participants noted, there was noticeably more latency in the spatial audio condition. The plugin used to convert B-format to binaural audio was chosen for its ease of use and good perceived sound clarity Although they might have lower degrees of fidelity, there are at least eight other alternatives that serve the same purpose but have lower processing latency [73]. As this latency is causing difficulties for playing and reducing the presence that the participants are experiencing, it is crucial that work be carried out to reduce it.

To guide future improvements in our experimental setup, we decided to conduct a comprehensive analysis of the latency contributions from each component in the audio processing and head tracking system. While sound processing is a critical factor, it is not the sole contributor to perceived delays. Other elements, such as the time it takes for a head movement to result in an audible change in sound, can significantly impact the user's sense of presence and immersion.

This examination will help us understand how each component influences the overall user experience and where improvements can be made to reduce the cumulative delay. For instance, enhancing the responsiveness of the system to user movements can create a more seamless and natural interaction or using plugins that have lower sound processing overhead can make the system feel more responsive, thereby enhancing the feeling of presence.

In Chapter 5, we will explore the methodologies used to measure these delays, carry out the measurements, present our results, and discuss the implications of our findings.

Addressing Visual Communication

The other major point raised by the participants was the lack of visual communication. As we saw, vision is used extensively in musical communication so we should find ways to enable such a communication channel. As introduced in Section 2.1, there have been demonstrations of low-latency video communication specifically geared towards musicians, so we know this is feasible.

Miscellaneous

Smaller pain points were brought up during this experiment, which might have been more participant-dependent. In particular, the headphones being wired got in the way of a participant, and another one mentioned feeling tied down and discouraged to move due to this. Perhaps the use of wireless headphones could be studied.

In addition, the binauralizer supports third party HRTF, so their personalization could be a long term goal for this project. Extensive research has been conducted on this topic, particularly in recent years. Several companies have already developed methods to personalize HRTFs using smartphones [74, 75]. This has the potential to be very beneficial for people for whom the 'average' HRTFs do not work well.

Finally, networked tests should be conducted, as this experiment was performed locally while testing the audio effects. Previous research, as discussed in Section 2.1, has demonstrated the feasibility of low-latency audio transmission over networks, and tools are available for such applications.

Chapter 5

Latency Characterization

Contribution of Authors

The measurement methodologies were devised by the author with feedback from Julien Boissinot, though inspired by Meyer-Kahlen [76]. Prof. Cooperstock proof-read, commented on, and helped to improve this chapter. The analysis and the discussion in this chapter is the work of the author.

5.1 Measurement Methodology

With many participants noticing or feeling impacted by the increased latency in the spatial audio condition, we needed to know the contributions of each element to the latency to inform future improvements. Figure 5.1 shows the complete block diagram of all components involved in the sound acquisition and processing pipeline. Two branches can be identified: the audio branch, from the microphone to the headphones for playback, and the IMU branch, from the movement measurements to the effects those movements have on the output sound.

The IMU we used is the BNO055 MEMS from Bosch, a 9-axis, absolute orientation sensor. It is sampled at 100 Hz by a Raspberry Pi Zero W running a Python script on Raspberry Pi OS. The host computer is a Lenovo machine with an AMD Threadripper Pro

5945WX 12-core processor, an Nvidia RTX 3080 Ti graphics card, and 32 GB of DDR4 RAM running at 3200 MHz. The audio interface in use is the RME Fireface UC, which connects to the host computer via USB and utilizes Audio Stream Input/Output (ASIO) drivers to communicate with Reaper, our DAW of choice.

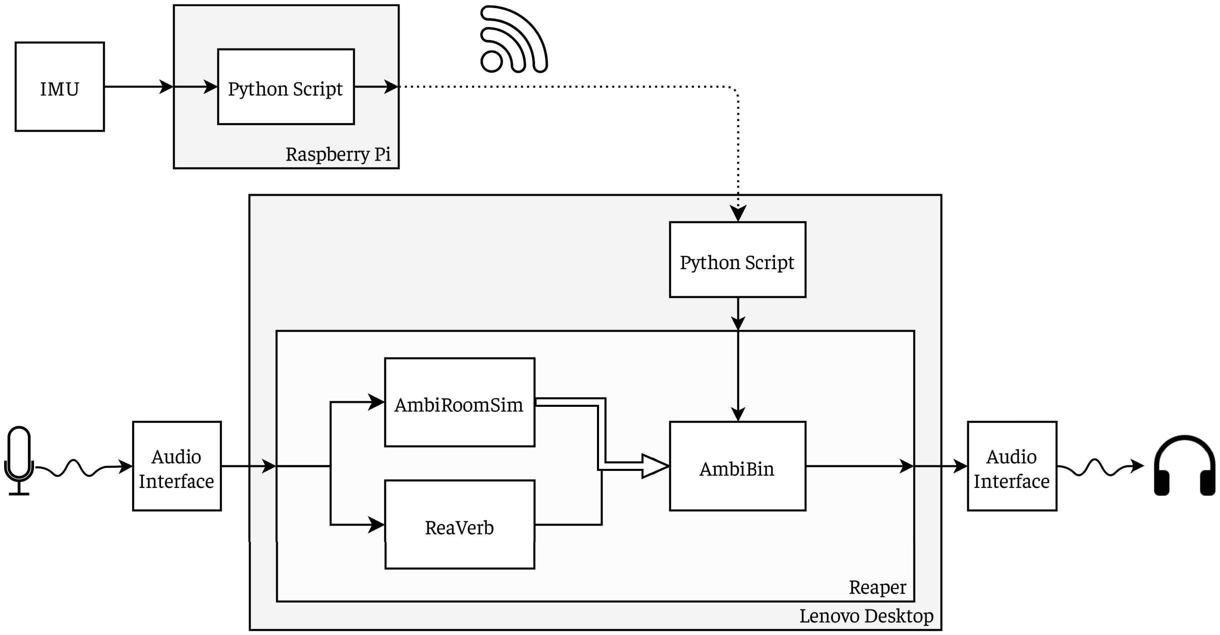

Figure 5.1 The sound processing pipeline for the experimental setup. Sound from the first participant is acquired through a microphone connected to the RME Fireface UC audio interface. Using ASIO drivers, Reaper retrieves the mono channel, which is subsequently processed through the AmbiRoomSim and ReaVerb plugins. The former transforms this input into a third-order Ambisonic signal and the latter adds the late reverberation. The output from the ReaVerb plugin is combined with the W channel (the omni-directional channel) of the Ambisonic signal, and the resulting mix is then routed into the AmbiBIN plugin. This converts the signal into a stereo, binaural signal given a head orientation measured by the IMU. The audio signal is finally sent to the audio interface, which plays it back through the headphones worn by the other participant.

5.1.1 The Audio Processing Branch

End-to-end audio delay

To start, the audio end-to-end delay is measured. This is the latency between sound being picked up by the microphone until is is played back by the headphones. Using an Agilent Technologies MSO6054A oscilloscope, the author measured the input and output audio signals by connecting to the wires going into and coming out of the audio interface. Using a secondary computer, a pure sinusoidal sound was played back through the headphone jack where a passive audio splitter was used to connect to the input of the interface and to the oscilloscope simultaneously. The output of the audio interface was provided as the second input channel to the oscilloscope. This way, we can observe the start of the audio signal before and after the processing by the experimental setup. We conducted measurements under various conditions: mirroring the user experiment with spatial audio and room acoustics (as depicted in Figure 5.1), replicating the control condition (direct playback without plugins), using only the AmbiRoomSim plugin (playing back the W channel), using only the AmbiBIN plugin, and using only the ReaVerb plugin.

ASIO drivers were used for the audio interface, as recommended by both Reaper and RME to minimize latency. Reaper and the audio interface were set at a 48 kHz sampling rate, which is standard in music applications, and used a block (or buffer) size of 128 samples, the minimum for the SPARTA suite [49] of plugins to work properly according to our tests, as is corroborated by Tomasetti [73].

5.1.2 The IMU Branch

End-to-end delay

For the IMU branch, the end-to-end delay is the time it takes for a movement to affect the audio signal. In the experiment, this represents the duration between someone moving their head and the ‘rotated’ audio signal being played back through the headphones.

The IMU data was multiplied by a factor of 3 such that a small movement would lead to a large 'instruction' for the AmbiBIN plugin, making the resulting change on the output waveform easier to measure. After being placed in a resting position, the IMU was hit with a conductive metal stick wired to one probe of the oscilloscope whose ground was connected to the 5 V power supply for the IMU. The hit causes a rotation of the device and a voltage drop is measured the moment the stick hits it. The second probe is attached to the output sound signal. The latency between the stick hitting and setting the IMU in motion and the change of the audio signal can therefore be measured.

Movement to data reception delay

We now wish to measure the delay between a movement occurring and the corresponding IMU data being received. To measure the start of the movement, a microphone on a microphone arm is placed a few centimeters away from and pointed towards the IMU at rest on a flat surface. When the stick hits the IMU, the microphone will pick it up and record the sound of the impact onto a track. Next, we use a Reaper plugin called TrackerTest [76] to receive the IMU data and convert the normalized values into waveforms for each degree of freedom (DoF). After recording the microphone and the TrackerTest tracks simultaneously, we can open the recordings using Audacity [77], a free audio editing software. We can then analyze the waveforms and measure the latency between multiple occurrences of the initial hit and the reception of the IMU movement data.

Input data to audio change

Once the AmbiBIN plugin receives the IMU data, there will be a delay before the output audio signal is affected. This is in part due to the convolution with the HRTF. However, there are other factors as well that vary depending on the plugin that is used [73]. To measure the incoming IMU data, we used the TrackerTest plugin again. The IMU output data was conditioned to round all measurements to either 90° or $-90°$, simulating an instantaneous rotation of the head by 180°. The room simulation was set up so that the virtual source was to the side of the position of the listener. This way, the audio signal output will shift most of its intensity from one ear to the other when an orientation change occurs. By recording the output signal to a track, we can again measure the delay between the IMU data change and the output track change. After the recording has been performed multiple times, the tracks are opened with Audacity. We can then measure the delay between each change and obtain a median value.

5.2 Results

Latency Measurements			
Type	Description	Latency (ms)	Standard Deviation (ms)
Audio end-to-end	with all plugins	42.43	
	with no plugins	7.80	
	with AmbiRoomSim	13.16	
	with ReaVerb	7.81	
	with AmbiBIN	39.82	
IMU	movement to Reaper	58	11.33
	Reaper to sound	9	1.26
	movement to sound	76.70	11.28

Table 5.1 A summary of all the latency measurements.

Table 5.1 displays a summary of all the latency measurements that were made. For the audio section, only three to five measurements were taken per category, as the latency exhibited no jitter and varied by mere fractions of a millisecond, attributable to measurement error. For the IMU section, 30 measurements were made for the 'movement to sound' category and 50 measurements were made for the two other categories since the jitter is very noticeable, likely due to the Wi-Fi connection between the Raspberry Pi and the main computer. The distribution of this data can be seen in Figure 5.2. The implications and explanations of those results will be discussed in the next chapter.

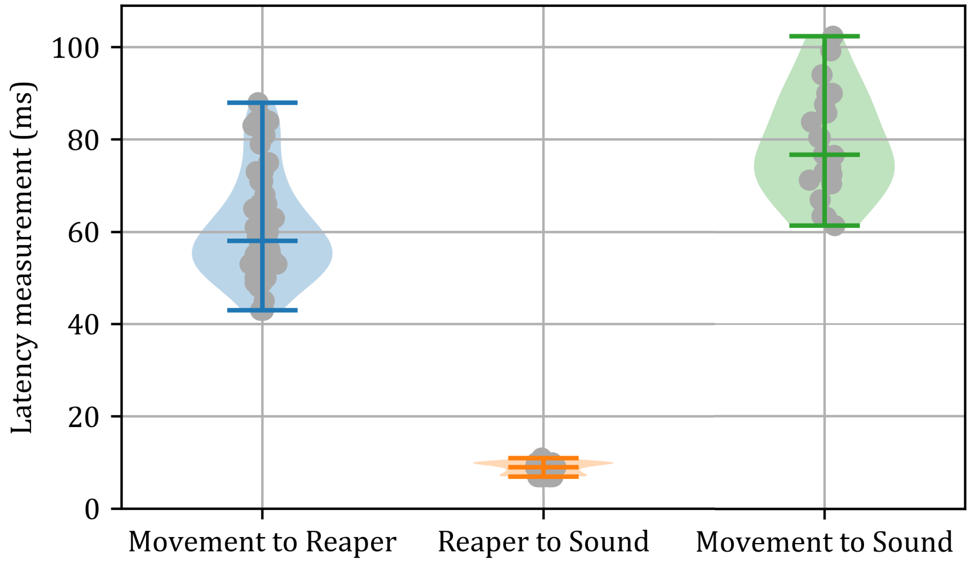

Figure 5.2 The distribution of the measurements for the three categories in the IMU branch. The top bar reflects the maximum, the bottom bar the minimum, and the middle bar the median of the data points. The 'movement to Reaper is the delay between a movement occurring and Reaper receiving the movement data. The 'Reaper to sound' category is the delay between Reaper receiving movement data and the sound adjusting as a consequence of that movement. The 'movement to sound' category is the delay between a physical movement occurring and the output audio signal being affected as a consequence of that movement.

5.3 Discussion

5.3.1 Factors in the Audio Delays

The end-to-end delay, with all plugins activated as in the user experiment, measures 42.43 ms. This is greater than the maximum of 30 ms that is recommended before slowdowns start occurring when the musicians play [10]. However, with no plugins and Reaper playing back sound immediately as it comes in, the end-to-end delay is only 7.80 ms. This number will be useful in the determination of individual contributions to the latency.

The end-to-end delay measured when only the AmbiRoomSim plugin was present and activated in the audio pipeline was 13.16 ms. By subtracting the latency with no plugins from it, we can obtain the individual contribution of AmbiRoomSim:

$$13.16 - 7.80 = 5.36\,\text{ms}$$

Similarly, the latency added by the AmbiBIN plugin is:

$$39.82 - 7.80 = 32.02\,\text{ms}$$

The latency measurements with only the ReaVerb (reverb) plugin active suggest that it either introduces no additional latency or that any increase is within the margin of error for our measurements. Since it is being used in parallel and not in series with the sound pipeline, this measurement confirms that it does not introduce delay onto the bypass. It could be relevant to make sure that the audio that is produced by this plugin 'overlaps' in the right location with the output of the other branch if both do not generate the same amount of delay.

With these results, we can now verify the coherence of our findings. We can add up the individual latencies to see if we find the end-to-end delay with every plugin activated.

$$\text{AmbiRoomSim} + \text{AmbiBIN} + \text{End-to-End w/o plugins} = \text{Total Latency}$$

$$5.36 + 32.02 + 7.80 = 45.18\,\text{ms}$$

Our result is 2.75 ms off from the end-to-end measurement which is a 6.5% error and can therefore be attributed to measurement inaccuracies. It is also possible that, when used in tandem, the buffers used by the plugins can overlap, increasing the efficiency slightly. These results also come close to the findings by Tomasetti [73] where AmbiRoomSim is reported to use 251 samples (5.23 ms at 48 kHz) and AmbiBIN uses 1530 samples (31.88 ms). This component contributes the majority of the latency within the audio chain. However, most other plugins reviewed in [73] introduce less latency, with some adding as little as 37 samples for room simulation and 16 samples for the binaural decoder. Therefore, we will need to see if these meet our needs for the future of the project. A visualization of the latencies can be seen in Figure 5.3

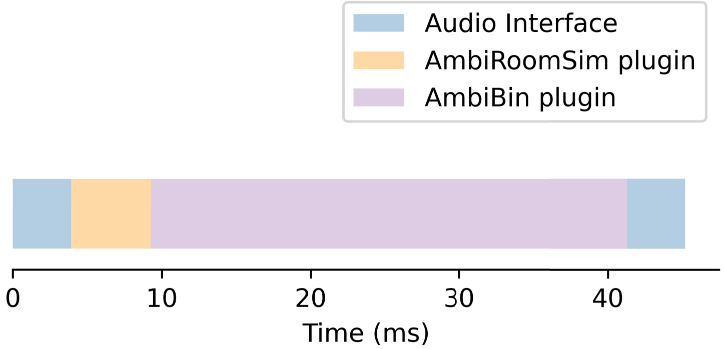

Figure 5.3 A summary and visualization of the additive delays in the audio processing pipeline. This assumes an even split between audio acquisition and playback by the audio interface.

5.3.2 Factors in the Movement to Sound Delays

The latency between a movement and its impact on sound is crucial for maintaining a sense of presence. To accurately simulate the changes in sound perception that occur when rotating the head, the audio processing must be adjusted in real-time. The minimum detectable latency is approximately 80 ms for the majority of people, and about 60 ms for the best listeners [78]. Past this threshold, users start losing some localization capabilities.

Measuring the time from an impulsive movement until the waveform adaptation yielded a median delay of 76.70 ms with a standard deviation of 11.28 ms, indicating the presence of jitter. Therefore, this delay is just under the threshold for noticeability by the average participant most of the time. Since our test population is musicians, it is likely that they would be classified as ‘good listeners’ and their threshold would be lower, potentially making the latency noticeable to them. Most of this jitter is due to the wireless connection, as will be discussed shortly.

The median delay between a movement and the reception of the data by Reaper was measured to be 58 ms with a standard deviation of 11.33 ms. Multiple factors could be at play here, from the IMU having an inherent latency between the movement and the measurement being available to be fetched over I2C, reading from the IMU register, the latency of the processing by the Raspberry Pi, and the network latency. The sampling rate of the BNO055 IMU is 100 Hz, or every 10 ms.

The median delay between the reception of an orientation vector and the sound output being affected accordingly was measured to be 9 ms with a standard deviation of 1.26 ms. This is surprising since the AmbiBIN plugin introduces an audio latency of over 30 ms but somehow reacts to data within 10 ms. Our hypothesis for this is that the plugin might be decoding from Ambisonic to stereo during the first 20 ms, and then transforms into binaural. The second stage would be when the head tracking data comes into play to rotate the decoded signal, a step that might be taking under 10 ms. This theory is reinforced by the fact that the HRIRs used by AmbiBin are 256 samples long, which leads to a theoretical $256/48000 = 0.005\overline{3} \approx 5.3$ ms of latency. The low standard deviation could be attributed to measurement error. The similarity in standard deviations for both the end-to-end and movement to Reaper measurements suggests that the jitter from the end-to-end latency is inherited by the movement to Reaper segment. This leads to the hypothesis that the wireless transmission over Wi-Fi could be to blame. Operating on a local and dedicated network could help reduce some of this latency and jitter. In his experiments, Wang [79] was able to measure latencies as low as 7.5 ms. The latency of the transmission between an ESP32 Wi-Fi microcontroller and a computer was measured, with the microcontroller sending 12 B OSC messages via a wireless access point. To obtain the lowest latencies, the access point had no encryption, the least busy Wi-Fi channel was selected, and measurements were carried out on a weekend to ensure low network traffic. He increased the transmission rate from 50 Hz to 2300 Hz and obtained latencies ranging from 6.76 ms to 10.67 ms. Notably, all transmission rates below 1000 Hz had a mean latency under 10 ms.

Network congestion was subsequently explored by measuring transmission latency at 100 Hz, with the number of additional devices ranging from zero to twelve. The mean transmission latency increased from 6.42 ms in the absence of additional devices to 16.45 ms with twelve devices present. For configurations with up to six supplementary devices, the mean latency remained below 10 ms.

We can sum up the individual latencies again to confirm their additive properties. Since the movement to Reaper and Reaper to sound delays were measured with a microphone, they also incorporate the audio acquisition delays, thereby reducing the measured latency. We can account for it by adding the delay back in.

$$\text{Movement to Reaper} + \text{Reaper to sound} + \text{Audio without plugins} = \text{Movement to sound}$$

$$58 + 9 + 7.80 = 74.8\,\text{ms}$$

Once again, we find a result close to our end-to-end measurement, 2.08 ms and 2.7% off from our expected value of 76.7 ms. It is unclear what the IMU latency is since it is not only measuring but it also computes the absolute orientation. The network, while not very congested, might have added significant latency due to both the Raspberry Pi and the host computer being on different sub networks managed by the university. While the Raspberry Pi transmitted over Wi-Fi, the computer was wired through an Ethernet port.

Chapter 6

Conclusion

The objective of this thesis was to attempt to create a system that will satisfy the project requirements as laid out in Section 2.1.1 and to evaluate the performance of the design decision on the target population: musicians. While a first method was explored, it quickly proved to be inadequate, and the Ambisonics-based solution arose to be a clear better performer while giving a clear path for improvement and integration into future iterations of the system. Following the design of a user experiment with two musicians playing together from separate rooms, we were able to evaluate the qualities and drawbacks of our solution. The binaural audio with room acoustics benefits participant engagement, immersion, and connectedness to their partner. However, the current design added an extra 34.5 ms of latency which the musicians noticed and were affected by. Participants also missed the ability to communicate visually as they normally do since our design did not feature a way for them to see each other. As our intention was to evaluate the audio performance of the system, we can say that the spatial audio with room acoustics experience is showing promising results, but the latency will need to be addressed before its performance can be considered to be acceptable.

Appendix A

User Study Questionnaire

A.1 Pre-Study Questionnaire

1. How old are you?
2. What instrument(s) do you play?
3. How long have you been playing?
4. Is the spatialization of sound important to you?

A.2 Post-condition Questions

1. What is something that you liked about what you just experienced?
2. What is something that you disliked about what you just experienced?

A.3 Post-Study Questions

1. Tell me about what you just experienced.
2. How did you feel?

3. Could you tell what the difference was? What were the biggest/ two biggest differences for you in the two experiences?
4. How did it impact you and your playing?
   (a) How did you make use of the spatial audio?
      i. If not, why?
   (b) How does the feeling of having the other person in the same room change the way you play? To what degree did you actually feel like the other person was in the room?
   (c) Do you engage more with the other musician?
5. Did you like the experience?
   (a) What is one thing you liked about the experience?
   (b) And one you didn't enjoy?
6. Which of these was your favorite experience and why?
7. Could you see this being a useful addition to a remote playing session?
8. What was missing from the audio experience?
9. What did you feel was a barrier to a better experience?
10. If you could change one thing about it, what would it be?
11. Can you identify any insights or realizations that emerged during or after the experiment?
12. Who did you feel was leading when playing?
13. Anything else you would like to share?

Bibliography

[1] J. V. Draper, D. B. Kaber, and J. M. Usher, “Telepresence,” *Human Factors*, vol. 40, pp. 354–375, Sept. 1998. Publisher: SAGE Publications Inc.

[2] R. Held, “Telepresence,” *The Journal of the Acoustical Society of America*, vol. 92, pp. 2458–2458, Oct. 1992.

[3] T. Letowski, “Sound quality assessment: concepts and criteria,” in *Audio Engineering Society Convention 87*, Audio Engineering Society, Oct. 1989.

[4] L. Kozma-Spytek and C. Vogler, “Factors affecting the accessibility of voice telephony for people with hearing loss: Audio encoding, network impairments, video and environmental noise,” *ACM Transactions on Accessible Computing*, vol. 14, pp. 1–35, Dec. 2021.

[5] Y. Qi and R. E. Hillman, “Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals,” *The Journal of the Acoustical Society of America*, vol. 102, pp. 537–543, July 1997.

[6] M. C. Killion and H. G. Mueller, “Twenty years later: A new count-the-dots method,” *The Hearing Journal*, vol. 63, p. 10, Jan. 2010.

[7] D. Botteldooren and B. D. Coensel, “Quality labels for the quiet rural soundscape,” in *Proceedings of the 35th International Congress and Exposition on Noise Control Engineering (Inter-Noise 2006)*, 2006.

[8] I. ISO, “12913-1: 2014: Acoustics—soundscape part 1: Definition and conceptual framework,” *ISO: Geneva, Switzerland*, 2014.

[9] M. Narbutt, S. O’Leary, A. Allen, J. Skoglund, and A. Hines, “Streaming VR for immersion: Quality aspects of compressed spatial audio,” in *2017 23rd International Conference on Virtual System & Multimedia (VSMM)*, pp. 1–6, Oct. 2017. ISSN: 2474-1485.

[10] N. Schuett, “The effects of latency on ensemble performance,” *Bachelor Thesis, CCRMA Department of Music, Stanford University*, 2002.

[11] A. Xu, W. Woszczyk, Z. Settel, B. Pennycook, R. Rowe, P. Galanter, J. Bary, G. Martin, J. Corey, and J. R. Cooperstock, "Real-time streaming of multichannel audio data over internet," *Journal of the Audio Engineering Society*, vol. 48, pp. 627–641, July 2000.

[12] J. R. Cooperstock, J. Roston, and W. Woszczyk, "Broadband networked audio: Entering the era of multisensory data distribution," in *18th International Congress on Acoustics*, (Kyoto, Japan), Apr. 2004.

[13] J.-P. Cáceres and C. Chafe, "JackTrip: Under the hood of an engine for network audio," in *Proceedings of International Computer Music Conference*, (San Francisco, California), pp. 509–512, International Computer Music Association, 2009. Backup Publisher: International Computer Music Association.

[14] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, pp. 1648–1661, Mar. 1992.

[15] R. M. Warren, *Auditory Perception: A New Synthesis*. Elsevier, Oct. 2013.

[16] T. Francart and J. Wouters, "Perception of across-frequency interaural level differences," *The Journal of the Acoustical Society of America*, vol. 122, pp. 2826–2831, Nov. 2007.

[17] D. Zotkin, J. Hwang, R. Duraiswaini, and L. Davis, "HRTF personalization using anthropometric measurements," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, pp. 157–160, Oct. 2003.

[18] F. L. Wightman and D. J. Kistler, "Sound localization," in *Human Psychophysics* (W. A. Yost, A. N. Popper, and R. R. Fay, eds.), pp. 155–192, New York, NY: Springer, 1993.

[19] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco, "Tori of confusion: Binaural localization cues for sources within reach of a listener," *The Journal of the Acoustical Society of America*, vol. 107, pp. 1627–1636, Mar. 2000.

[20] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *ACTA ACUSTICA UNITED WITH ACUSTICA*, vol. 91, 2005.

[21] P. D. Coleman, "An analysis of cues to auditory depth perception in free space," *Psychological Bulletin*, vol. 60, pp. 302–315, May 1963.

[22] M. Talbot-Smith, "8 - Sound, speech and hearing," in *Telecommunications Engineer's Reference Book* (F. Mazda, ed.), pp. 8–1, Butterworth-Heinemann, Jan. 1993. Num Pages: 8-16.

[23] A. S. Edwards, "Accuracy of auditory depth perception," *The Journal of General Psychology*, vol. 52, pp. 327–329, Apr. 1955. Publisher: Routledge _eprint: https://doi.org/10.1080/00221309.1955.9920247.

[24] E. A. Gamble, "Minor studies from the psychological laboratory of Wellesley College: Intensity as a criterion in estimating the distance of sounds," *Psychological Review*, vol. 16, no. 6, pp. 416–426, 1909. Place: US Publisher: The Review Publishing Company.

[25] T. Z. Strybel and D. R. Perrott, "Discrimination of relative distance in the auditory modality: The success and failure of the loudness discrimination hypothesis," *The Journal of the Acoustical Society of America*, vol. 76, pp. 318–320, July 1984.

[26] W. E. Simpson and L. D. Stanton, "Head movement does not facilitate perception of the distance of a source of sound," *The American Journal of Psychology*, vol. 86, no. 1, pp. 151–159, 1973. Publisher: University of Illinois Press.

[27] Z. Qian, D. Shang, Y. Hu, X. Xu, H. Zhao, and J. Zhai, "A Green's function for acoustic problems in Pekeris waveguide using a rigorous image source method," *Applied Sciences*, vol. 11, p. 2722, Jan. 2021.

[28] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pp. 99–102, Oct. 2001.

[29] M. D. Burkhard and R. M. Sachs, "Anthropometric manikin for acoustic research," *The Journal of the Acoustical Society of America*, vol. 58, pp. 214–222, July 1975.

[30] K. Inanaga, H. Kon, G. Rasmussen, P. Rasmussen, and Y. Riko, "Study and consideration on symmetrical KEMAR HATS conforming to IEC60959," in *Audio Engineering Society Convention 126*, Audio Engineering Society, May 2009.

[31] X. Tian, Z. Fu, and L. Xie, "An experimental comparison on KEMAR and BHead210 dummy heads for HRTF-based Virtual auditory on Chinese subjects," in *IET 3rd International Conference on Wireless, Mobile and Multimedia Networks (ICWMNN 2010)*, pp. 369–372, Sept. 2010.

[32] C. J. Chun, J. M. Moon, G. W. Lee, N. K. Kim, and H. K. Kim, "Deep neural network based HRTF personalization using anthropometric measurements," in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.

[33] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, 2002.

[34] D. Alais and D. Burr, “The ventriloquist effect results from near-optimal bimodal integration,” *Current Biology*, vol. 14, pp. 257–262, Feb. 2004. Publisher: Elsevier.

[35] U. Schlemmer, “Reverb design,” in *Pure Data Convention, Weimar-Berlin, Germany*, 2011.

[36] S. T. Neely and J. B. Allen, “Invertibility of a room impulse response,” *The Journal of the Acoustical Society of America*, vol. 66, pp. 165–169, July 1979.

[37] J. W. Cooley and J. W. Tukey, “An algorithm for the machine calculation of complex Fourier series,” *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, 1965. Publisher: American Mathematical Society.

[38] T. G. Stockham, “High-speed convolution and correlation,” in *Proceedings of the April 26-28, 1966, Spring joint computer conference*, (Boston, Massachusetts), p. 229, ACM Press, 1966.

[39] F. Wefers, *Partitioned convolution algorithms for real-time auralization*. No. Band 20 in Aachener Beiträge zur technischen Akustik, Berlin: Logos Verlag Berlin GmbH, 2015.

[40] O. Warusfel, “Listen HRTF database,” *online, IRCAM and AK, Available: http://recherche.ircam.fr/equipes/salles/listen/index.html*, 2003.

[41] A. V. Oppenheim and R. W. Schafer, *Discrete-time Signal Processing*. Pearson, 2010.

[42] S. Hameed, J. Pakarinen, K. Valde, and V. Pulkki, “Psychoacoustic cues in room size perception,” *Journal of the Audio Engineering Society*, May 2004.

[43] J. S. Abel, N. J. Bryan, P. P. Huang, M. Kolar, and B. V. Pentcheva, “Estimating room impulse responses from recorded balloon pops,” in *Audio Engineering Society Convention*, vol. 129, Nov. 2010.

[44] E. A. Lehmann and A. M. Johansson, “Prediction of energy decay in room impulse responses simulated with an image-source model,” *The Journal of the Acoustical Society of America*, vol. 124, pp. 269–277, July 2008.

[45] L. Kelley, D. D. Carlo, A. A. Nugraha, M. Fontaine, Y. Bando, and K. Yoshii, “RIR-in-a-box: Estimating room acoustics from 3d mesh data through shoebox approximation,” in *INTERSPEECH*, (Kos International Convention Center, Kos Island, Greece), Sept. 2024.

[46] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, “AmbiX-a suggested ambisonics format,” in *Ambisonics Symposium*, vol. 2011, June 2011.

[47] J. Ahrens, *Analytic methods of sound field synthesis*. T-labs series in telecommunication services, Berlin: Springer, 2012.

[48] "REAPER | Audio Production Without Limits."

[49] L. McCormack and A. Politis, "SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *2019 AES International Conference on Immersive and Interactive Audio*, Audio Engineering Society, Mar. 2019.

[50] A. Neidhardt, C. Schneiderwind, and F. Klein, "Perceptual matching of room acoustics for auditory augmented reality in small rooms - Literature review and theoretical framework," *Trends in Hearing*, vol. 26, Jan. 2022.

[51] T. I. Niaounakis and W. J. Davies, "Perception of reverberation time in small listening rooms," *Journal of the Audio Engineering Society*, vol. 50, pp. 343–350, May 2002.

[52] M. Vigeant and R. Celmer, "Effect of experimental design on the results of clarity-index just-noticeable-difference listening tests," in *20th International Congress on Acoustics 2010, ICA 2010 - Incorporating Proceedings of the 2010 Annual Conference of the Australian Acoustical Society*, June 2010.

[53] F. del Solar Dorrego and M. C. Vigeant, "A study of the just noticeable difference of early decay time for symphonic halls," *The Journal of the Acoustical Society of America*, vol. 151, pp. 80–94, Jan. 2022.

[54] H. Steffens, S. van de Par, and S. D. Ewert, "The role of early and late reflections on perception of source orientation," *The Journal of the Acoustical Society of America*, vol. 149, pp. 2255–2269, Apr. 2021.

[55] H. Yan, Q. Shan, and Y. Furukawa, "RIDI: Robust IMU double integration," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 621–636, Sept. 2018.

[56] M. Nazarahari and H. Rouhani, "Sensor fusion algorithms for orientation tracking via magnetic and inertial measurement units: An experimental comparison survey," *Information Fusion*, vol. 76, pp. 8–23, Dec. 2021.

[57] S. O. H. Madgwick, A. J. L. Harrison, and R. Vaidyanathan, "Estimation of IMU and MARG orientation using a gradient descent algorithm," in *2011 IEEE International Conference on Rehabilitation Robotics*, pp. 1–7, June 2011. ISSN: 1945-7901.

[58] Android, "Motion sensors | Sensors and location."

[59] Apple, "Getting processed device-motion data."

[60] J. A. Olson, D. A. Sandra, E. S. Colucci, A. Al Bikaii, D. Chmoulevitch, J. Nahas, A. Raz, and S. P. L. Veissière, "Smartphone addiction is increasing across the world: A meta-analysis of 24 countries," *Computers in Human Behavior*, vol. 129, p. 107138, Apr. 2022.

[61] K. Jambrosic, M. Krhen, M. Horvat, and T. Jagust, "Measurement of IMU sensor quality used for head tracking in auralization systems," in *Forum Acusticum*, (Lyon, France), pp. 2063–2070, Dec. 2020.

[62] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, pp. 77–101, Jan. 2006.

[63] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-accurate speech transcription of long-form audio," *INTERSPEECH 2023*, 2023.

[64] R. Rampin and V. Rampin, "Taguette: open-source qualitative data analysis," *Journal of Open Source Software*, vol. 6, no. 68, p. 3522, 2021.

[65] L. M. Given, *The Sage encyclopedia of qualitative research methods*. Los Angeles (Calif.): Sage, 2008.

[66] T. X. Barber and M. J. Silver, "Fact, fiction, and the experimenter bias effect," *Psychological Bulletin*, vol. 70, no. 6, Pt.2, pp. 1–29, 1968. Place: US Publisher: American Psychological Association.

[67] V. Y. Oviedo, K. A. Johnson, M. Huberth, and W. O. Brimijoin, "Social connectedness in spatial audio calling contexts," *Computers in Human Behavior Reports*, vol. 15, p. 100451, Aug. 2024.

[68] T. Potter, Z. Cvetković, and E. De Sena, "On the relative importance of visual and spatial audio rendering on VR immersion," *Frontiers in Signal Processing*, vol. 2, Sept. 2022. Publisher: Frontiers.

[69] P. Larsson, D. Västfjäll, and M. Kleiner, "Effects of auditory information consistency and room acoustic cues on presence in virtual environments," *Acoustical Science and Technology*, vol. 29, no. 2, pp. 191–194, 2008.

[70] F. Seddon and M. Biasutti, "A comparison of modes of communication between members of a string quartet and a jazz sextet," *Psychology of Music*, vol. 37, pp. 395–415, Oct. 2009. Publisher: SAGE Publications Ltd.

[71] F. A. Seddon, "Modes of communication during jazz improvisation," *British Journal of Music Education*, vol. 22, pp. 47–61, Mar. 2005.

[72] M. Iorwerth and D. Knox, “Playing together, apart,” *Music Perception*, vol. 36, pp. 289–299, Feb. 2019.

[73] M. Tomasetti, A. Farina, and L. Turchet, “Latency of spatial audio plugins: a comparative study,” in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, (Bologna, Italy), pp. 1–10, IEEE, Sept. 2023.

[74] B. Zhi, D. N. Zotkin, and R. Duraiswami, “Towards fast and convenient end-to-end HRTF personalization,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 441–445, May 2022. ISSN: 2379-190X.

[75] C. Guezenoc and R. Seguier, “HRTF individualization: A survey,” in *145th Audio Engineering Society Convention*, 2018.

[76] N. Meyer-Kahlen, M. Kastemaa, S. J. Schlecht, and T. Lokki, “Measuring motion-to-sound latency in virtual acoustic rendering systems,” *Journal of the Audio Engineering Society*, vol. 71, no. 6, pp. 390–398, 2023. Publisher: Audio Engineering Society.

[77] “Audacity ® | Free Audio editor, recorder, music making and more!.”

[78] D. S. Brungart, B. D. Simpson, and A. J. Kordik, “The detectability of headtracker latency in virtual audio displays,” *International Conference on Auditory Display*, July 2005.

[79] J. Y. Wang, *Analysis of Wireless Interface Latency and Usability for Digital Musical Instruments*. PhD thesis, McGill University, 2021. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2024-12-11.