Royal Holloway, University of London

Doctoral Thesis

The Reality of Virtual Voice Production in Performance Spaces

Author:

Florence Roberts

Supervisor:

Professor David Howard

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Biosignals and Intelligent Systems Group
Electronic Engineering

June 2023

Abstract

Virtual reality opens the door to creating experiences that can take the user wherever or whenever they want, potentially even to places that are inaccessible to many people. Over the last decade, there has been a sharp increase in the popularity and availability of commercial VR headsets for both the worlds of entertainment and research. This has led to a revived interest in improving audio immersion in virtual environments, as a way to establish greater immersion overall.

This thesis provides an overview into various methods of auralisation for virtual spaces, investigating our ability to firstly recreate real acoustic environments naturally, and secondly to discover how true these models need to be to their corresponding real environment to maintain immersion, and what specific cues contribute most to a person's ability to feel fully immersed. Results show that for spaces in which a person is only listening to audio cues, an approximation of a real aural environment is sufficient to immerse a participant, showing a slight tendency of preference towards the falsified model in the case where a user was free to move about in a reverberant acoustic environment. Furthermore, when presented with the opportunity to freely manipulate reverberant tail length and first reflection delay in real-time for their own singing voices, immersion was maintained for users even when results differed slightly from the real measured values.

For Dan and Siobhan, who have been with me since the beginning. Undergraduate, masters, and now PhD: we've only gone and done it!

And in memory of Joshua Bolton (2000 - 2020), who showed to me just what it meant to never give up, even in the face of unimaginable difficulties. Here's to you Joshibi.

Acknowledgements

Firstly, a massive thank you goes to my supervisor, Professor David Howard. Thank you for showing me that I can continue to combine my love for science and music. Thank you for the *should-be-a-twenty-minute-update-meeting-that-turns-into-forty-extra-minutes-of-just-chatting* meetings. And thank you for being so supportive during the hardest year of my life.

Thank you to everyone who participated as singers, speakers, and listeners for my hypotheses. I literally would not have results or a PhD without you. Extra special shout out to the Gamesoc and HvZ community members who were always willing to participate in my research, and to the Jane Holloway Choir for being recorded for an experiment.

To all my actual and adopted housemates past and present. We may have been forcibly trapped together for the last couple of years, but if I had to be stuck in a lockdown once more, I would do it all again with you. Lunches in the garden, paddling pool in the garden, board games in the garden, cocktails in the garden, BBQs in the garden... Man, I should really also acknowledge our garden...

To the organ scholars at Royal Holloway - I once again apologise for the amount of times I popped balloons in the chapel. I really appreciate the patience you granted me.

For those who have helped and supported me throughout these last 4 years. Siobhan, we've got through our undergrad, our masters, and our PhDs together, and I am so glad to be able to call you my friend. Aletheia, if it were not for the pandemic, I probably would

never have met you. And that would have been a massive shame. Thank you for being my accountability buddy, and I am forever grateful that you messaged me after that zoom breakout room.

To my closest friends: there are too many of you talk about, but you hopefully know who you are. It is one of the blessings of having been at this university for forever. Thank you for sticking around in my life for so long and for providing endless joy, fun, and gaming. To my families who, even when they have no idea what I am talking about, continue to be proud and support me from afar.

To Dan Woods. I could not have finished this without you, and your love and support in all things that I do. For keeping me grounded when I am unnecessarily stressed, to being my rubber duck when I am working things out, to uplifting me for the work I do day to day, whether it was an easy day or hard one, I thank you. And thank you for proof reading this thesis!

And last but not least, to myself. When I first set out to apply for a PhD, I was faced with scrutiny over my reasoning as to why I wanted to do one. For anyone who happens to have read this far, know you can set out and achieve whatever you want to do. Do not listen to those who want to pull you down, and if you want to achieve something, you will do it.

You've got this. I believe in you.

Declaration of Authorship

I declare that this thesis is a presentation of original work, and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All contributions from outside sources, through direct contact or publications, have been explicitly stated and referenced.

Contents

1 Introduction **1**

1.1 Immersion 2

1.1.1 Immersion as defined in other literature 3

1.1.2 Definition of immersion 4

1.2 Realism, naturalness and authenticity 6

1.3 Thesis aims 7

1.4 Hypothesis 7

1.5 Novelty of research 8

1.6 Structure of thesis 9

2 Theory and Literature **12**

2.1 Virtual Reality 13

2.1.1 Game engines 15

2.2 Virtual recreation of real environments 17

2.2.1 3D modelling 17

2.2.2 Camera-based modelling 17

2.2.3 LiDAR 17

2.2.4 360° Imagery and video 18

2.3 VR and immersive audio - development of sound systems 19
2.3.1 Head-Related Transfer Functions (HRTFs) 19
2.3.2 Spatial audio system development 20
2.3.3 Binaural 23
2.3.4 Ambisonics 25
2.4 Auralisation of real environments 28
2.5 Acoustic modelling: Impulse Responses 29
2.5.1 Composition of an Impulse Response 29
2.5.2 Room Impulse Response from well defined signals 31
2.5.3 Room Impulse Response from an excitation signal 32
2.6 Convolution of Room Impulse Responses 33
2.6.1 Direct form convolution 34
2.6.2 Frequency domain convolution 35
2.7 Acoustic modelling: computational 36
2.7.1 Geometrical acoustics: image source 37
2.7.2 Geometrical acoustics: ray tracing 39
2.7.3 Resonance Audio 41
2.7.4 Geometrical acoustics: beam tracing 42
2.8 Recreating physical environments aurally 43
2.8.1 Virtual immersive audio experiences 43
2.8.2 Virtual interactive immersive performances 46
2.8.3 Reverberation and perception 51
2.9 Summary 54

3 Effectiveness of Simple Audio Recreation 56

3.1 Real - world audio experiences... 58
3.1.1 Concept and motivation 60
3.2 Recording in real acoustic environments 61
3.2.1 Audio sample recording 61
3.2.2 Impulse Response measurement 65
3.2.3 Scenario recording set-up 65
3.3 Rendering 66
3.3.1 Audio rendering 66
3.3.2 Video rendering 67
3.4 Method 69
3.4.1 Set-up of virtual environments 69
3.4.2 Pilot study 69
3.4.3 Further study adaptations 70
3.4.4 Procedure 71
3.5 Results 72
3.5.1 Participants 72
3.5.2 Questionnaire 72
3.5.3 Questionnaire results 73
3.6 Discussion 78
3.6.1 Further exploration and improvements 84
3.7 Summary and conclusions 87
4 Measuring and Modelling Physical Spaces Virtually 89
4.1 Sound produced from acoustic modelling... 92
4.2 Creating the visual chapel model 93

4.2.1 CAD and modelling 95
4.3 Measurement and fabrication of RT60s 98
4.3.1 Measuring Impulse Response frequency bands 99
4.3.2 Modelling Impulse Response frequency bands in Resonance Audio . . 100
4.3.3 Implementation of the Room Impulse Responses 102
4.3.4 Implementation of the sound sources 103
4.3.5 Discovered issues and solutions 105
4.4 Method 107
4.4.1 Layout of virtual space 107
4.4.2 Procedure 109
4.5 Results 110
4.5.1 Participants 110
4.5.2 Questionnaire 111
4.5.3 Questionnaire results 112
4.6 Discussion 119
4.6.1 Immersion, interactivity and user experience 119
4.6.2 Audio-visual conflict 121
4.6.3 Reverberation time disparity 122
4.6.4 Extended exploration 124
4.6.5 Further research 127
4.7 Summary and conclusions 129

5 The Design and Implementation of a Real-Time Audio Convolution System 131

5.1 Concept for the real-time convolution system 133

5.1.1 Desired outcome 133

5.1.2 Real-time convolution 133

5.1.3 Real-time manipulation 134

5.1.4 Usability 134

5.2 Real-time singer voice feedback 134

5.2.1 Pure Data patch for real-time audio manipulation 135

5.2.2 Measuring the latency of a system 137

5.3 Bela and ultra-low latency 139

5.3.1 Bela and Pure Data 140

5.4 Real-time convolution with Bela in C++ 142

5.4.1 Convolver creation 143

5.5 Audio processing programme 145

5.5.1 Key differences from the Pure Data patch 146

5.5.2 Real-time audio processing 147

5.5.3 Circular buffers and pointers 148

5.5.4 Render function 150

5.5.5 Graphical User Interface 151

5.5.6 Physical interface conversion 152

5.6 Conclusion 155

6 Some Sound Cues Contribute More to... 156

6.1 Introduction 156

6.2 We will rock you 157

6.3 Singing in virtual performance spaces 160

6.3.1 Motivation 160

6.3.2 System overview 161

6.4 Recording 162

6.4.1 Audio capture 162

6.4.2 360° Video capture 162

6.4.3 Song choices and recording layouts 163

6.5 Rendering 165

6.5.1 Adobe Premiere Pro 165

6.5.2 REAPER and upmixing for loudspeaker setups 166

6.5.3 Bela Box 167

6.6 Installation and implementation 168

6.6.1 VR visuals setup 169

6.6.2 Choir audio setup 170

6.6.3 Real-time audio setup 170

6.6.4 Procedure 171

6.7 Results 172

6.7.1 Participants 172

6.7.2 Dial values 173

6.7.3 Questionnaire results 174

6.8 Discussion 180

6.8.1 Discussion: dial values 180

6.8.2 Discussion: questionnaire 182

6.8.3 Discussion: realism and immersion 184

6.8.4 Further exploration and improvements 185

6.9 Summary and conclusions 186

7 Conclusions 188

7.1 Summary of thesis 189

7.2 The hypothesis revisited 194

7.2.1 Novelty of research 196

7.3 Further work 198

7.3.1 Further development for real-time singing experiences 198

7.3.2 Improving the fully explorable virtual environments 200

7.4 Applications of research 201

7.4.1 Real-time convolution 201

7.4.2 Rehearsal and performance 201

7.4.3 Psychoacoustics 202

7.4.4 Virtual reality and immersive audio 202

7.4.5 Architectural design past and present 203

7.5 Final comments 203

A Instructions and Questionnaire for Study 1 (Chapter 3) 205

B Instructions and Questionnaire for Study 2 (Chapter 4) 216

C Instructions and Questionnaire for Study 3 (Chapter 5) 227

List of Figures

1.1 Summary of different immersion terms from the literature. 4

2.1 A diagram of a 5.1 surround sound system layout, where the arrow indicates 0 degrees. The blue figure in the centre of the image represents a listener, and the five grey boxes the speakers. 23

2.2 Symmetrical mazes for testing the two audio conditions. For each test, the participant would start in the centre [1]. 24

2.3 Soundfield Ambisonic Microphone capsules [2] and the names of each capsule. 26

2.4 Representation of a Room Impulse Response. 30

2.5 A block diagram for the Finite Impulse Response (FIR) filter implemented with a direct form convolution, where *x[n]* is the input, ▷ is amplification, z^{-1} is delay, $\bigoplus$ is summation, and *y[n]* the output. 34

2.6 A visual representation of frequency domain convolution. 36

2.7 Examples of the image source method for geometrically simulating a sonic environment. 38

2.8 A simplified diagram of early reflections in geometric acoustics [3]. 40

2.9 A simulation of how real sound waves travel for use in Resonance Audio, highlighting occlusion effects [4]. 41

2.10 Simplified virtual models of the Cathedral of Seville and the Royal Chapel. Geometrical model before simplifications (left); simplified visual model (centre); and simpli- fied acoustic model (right) [5]. 44

2.11 Photo of the Royal Chapel of the Cathedral of Seville (left) and the corresponding virtual VR model (right) [5]. 45

2.12 GoPro rig and Eigenmic set-up for singer replacement [6]. 47

2.13 The set-up for the virtual reality experience [6]. 48

3.1 Two spectrograms comparing the sample recording location's acoustics with an original clean recording of an orchestra. 63

3.2 A graph to show the frequency response (averaged and compensated) for the Anker Soundcore 2 Speaker [7]. 64

3.3 Part of the interface of Facebook 360 Spatial Workstation. Shown is the placement of a mono track (blue sphere) at the position of the visible speaker system in the recorded visuals for the listener's (camera) position. 67

3.4 A group of bar charts which show the number of responses for the live and falsified scenarios for the question *Of the two scenarios, which one did you think used the live recording of the location?* Participants were choosing between videos randomly labelled 1 and 2 and did not know which audio was falsified. 74

3.5 The results for the questions asking participants to choose which video they thought was the live recording and state their certainty in their answers for each pair of recordings. 75

3.6 A showcase of a selection of the responses and reasoning for the choices made when listening to both scenarios. 76

3.7 A group of bar charts to show the number of responses of *Yes*, *Not Sure*, and *No* for the question *Did the audio ever make you feel as though you were at the location?* for both the Chapel and meadow scenarios. 77

3.8 A graph to compare the amplitude of the frequencies created from passing white noise through the IR of the meadow with the average frequencies and higher harmonics of the female and male voices. 80

3.9 A pair of spectrograms comparing the live and falsified orchestral Chapel recordings. 82

3.10 (a) VRate's UI for adding a questionnaire into Unity [8]. (b) VRQuestionnaireToolkit's UI for Unity using NASA TLX [9]. 85

3.11 A scene created in Unreal Engine demonstrating the concept of utilising head-high spheres for a virtual body to walk into to combat the feeling of a Sky-sphere feeling infinitely far away from the user. 86

4.1 A photo of the Royal Holloway Chapel, 2021. 94

4.2 A dimensional drawing of the north side of Royal Holloway University of London's Founder's Building provided by the CAD Technical Support team. 95

4.3 A side by side comparison of a CAD drawing of the chapel with one of the original architectural drawings. This was utilised to find the radius of the arched ceiling. 96

4.4 The floor plan for the ground and first floor of Royal Holloway's chapel. . . . 97

4.5 Initial chapel modelling in ProBuilder. The different colours are just for easy identification of separate components of the model. 97

4.6 The model of Royal Holloway's Chapel made in Autodesk Inventor Professional. The figure on the right displays some of the detailing inside, which is where participants will virtually be while taking part in the study. 98

4.7 (a) Unity *RT60s for frequency bands* panel. (b) Source code for the frequency bands and their ranges. 101

4.8 Measured (Live Recording) and Fabricated (Resonance Audio) results of Reverberation Time. 102

4.9 Two figures showing from a birds-eye view, the two scenarios A and B. . . . 108

4.10 The results for the questions about choosing which acoustic model was made from real-world measurements. 113

4.11 The results for the questions asking about participants' feelings on which model was the most 'realistic' sounding for them. 116

5.1 A novel Pure Data patch that, in real-time, can alter the way a person's voice sounds through the manipulation of 3 sliders; IR length, Initial Time Delay Gap, and direct audio volume. 136

5.2 (a) A photo of the standard Bela Board with audio adapter cables attached for use with headphones and a microphone; (b) A pin diagram of the Bela board taken from the Bela local IDE [10]. 140

5.3 The Pure Data patch reworked for the Bela board that was ultimately discarded.142

5.4 A visual representation of how the impulse responses are split into blocks of varying sizes, where h is the impulse response and N is the block size for which the FFT convolution becomes more efficient than the direct form convolution. 143

5.5 A visual representation showing how latency is incurred in a system. Each block of audio is a buffer, and the pink highlighted buffers show why the latency of a system is at minimum twice the buffer length. 148

5.6 A visual representation of an inefficient buffer system where every stored sample is moved every time a new sample is added. 149

5.7 A visual representation of how circular buffers work. The blue arrow represents the write pointer, and the orange arrow the read pointer. 149

5.8 The Graphical User Interface (GUI) for a Bela patch that enables user interaction with the convolution algorithm in real-time. 151

5.9 A photo of the Bela Box alongside a circuit diagram of the internal hardware. 154

6.1 The rear 180° view from behind the virtual singer of the volunteer choir singing We Will Rock You, alongside the recording equipment setup used. 158

6.2 The positioning of the choir members and the recording set-up. The arrow represents the forward-facing orientation of the camera and microphone. . . . 159

6.3 The three capture positions during the recording of the song *Hymn to the Fallen*. The white, numbered circles represent the camera locations, the arrows the forward-facing orientation of the recording setup, and the rectangular box the conductor and pianist. Recording 1: soprano point of view. Recording 2: alto point of view. Recording 3: tenor and bass point of view. The accompanying photos show the forward-facing view from the camera. 164

6.4 The recording layout for the improvised group piece. The arrow represents the forward facing direction of the recording equipment and the view of this is shown in the accompanying photo. 165

6.5 The working window for the speaker layout controlled by Rapture3D software. 166

6.6 A diagram detailing the positions of all equipment used for the study across two rooms. 168

6.7 Photos detailing the layout of the experiment space. 169

6.8 Photos showing the positions of the off ear speakers and microphone and how they are connected to the headset and Bela Box. 171

6.9 A graph showing the spread of impulse responses selected by participants. *100%IR* is the actual impulse response for the chapel. 174

6.10 A set of bar graphs displaying the answers for participants who sang with the mixed unison choir singing *Silent Night*. 176

6.11 A graph showing the responses from members of the Jane Holloway Choir to the statements posed. 179

6.12 A graph showing the mean scores (and standard error) of singers' responses to the questionnaire 182

List of Tables

3.1 Chi-square values for each scenario comparison, where O_L represents number of votes for the live recording, O_F the number of votes for the falsified recording, and χ^2_{cv} the critical value for two-tailed, one degree of freedom chi-square test (3.841). 81

3.2 Table comparing type of headset to average correct identifications of live audio recording 83

4.1 Averaged measured results of Reverberation Time as defined by the standard ISO 3382-1. Anomalous values are indicated through red text. 100

4.2 Percentages of *Confident, Somewhat Confident, Neutral, Somewhat Unsure,* or *Unsure* responses for the questions on choosing which model was made from real-world data for both scenarios. 114

6.1 A table highlighting the spread of results for participants choosing what they thought the correct ITDG for the first reflection for the chapel was. 174

6.2 A table displaying respondents' choices of ITDG grouped into ±5 ms ranges. 181

Acronyms

VR Virtual Reality

XR Extended Reality

FIR Finite Impulse Response

RT60 Reverberation Time

IDE Integrated Development Environment

FFT Fast Fourier Transform

GUI Graphical User Interface

Pd Pure Data

RIR Room Impulse Response

IR Impulse Response

HMD Head Mounted Display

CNN Convolutional Neural Network

CAD Computer-Aided Design

LiDAR Light Detection and Ranging

HRTF Head-Related Transfer Function

ILD Interaural Level Difference

ITD Interaural Time Difference

MLS Maximum Length Sequence

BRIR Binaural Room Impulse Response

FER Forward Early Reflection

RER Reverse Early Reflection

LR Late Reverberation

VSS Virtual Singing Studio

AES Audio Engineering Society

FB360 Facebook 360 Spatial Workstation

EDT Early Decay Time

AR Augmented Reality

SATB Soprano, Alto, Tenor and Bass

ITDG Initial Time Delay Gap

Chapter 1

Introduction

> *“...All sounds belong to a continuous field of possibilities lying within the comprehensive dominion of music. Behold the new orchestra: the sonic universe! And the musicians: anyone and anything that sounds!”*
>
> - R. M. Schafer, *The Soundscape: Our Sonic Environment and the Tuning of the World* [11]

Virtual reality provides us with the wonderful opportunity to create experiences that, in the real world, could otherwise be inaccessible to many people. Although the current VR headsets were originally intended for gaming purposes, allowing the user to become more immersed in a fictional world, this technology has expanded into the research world more and more over the last decade (see Chapter 2), allowing for studies exploring immersive experiences to improve in quality. However, there is still more to be done to improve a person's immersion in a virtual space.

Over the last decade, there has been a resurgence in virtual reality technology as a whole. This has led to revived interest in improving audio immersion in video games, and in the world of research, as a means of establishing a greater sense of engrossment in virtual environments. Over the last couple of years in particular, a greater push to improve immersivity arose due to the COVID-19 pandemic, where people could not leave their houses or see people outside of those they lived with. For escaping to a more realistic feeling virtual environment or the desire for social interaction virtually, the pandemic spurred the adoption of VR into general life [12]. However even with this push, acoustic immersion systems still flags behind visual immersion systems. But what is meant by immersion?

1.1 Immersion

The fields of virtual reality and spatial audio have grown over the last decade, leading to ambiguity around terminology for describing improved visual and auditory experiences [13]. Despite attempts from multiple authors to define terminology related to VR environments and perception, particularly in terms of ‘realism’, ‘naturalness’, and ‘immersion’, there is still confusion regarding these concepts, which may vary depending on one’s perspective and academic discipline. This poses a challenge for collecting opinions on the subject, as well as defining them for the average consumer. For example, immersion is frequently used as a marketing term for commercial products, and the definition has become associated with product quality, in that a product being “more immersive” makes it better. These observations highlight the need for a consistent definition of these terms, which other researchers have attempted in the past [14, 15]. However, due to a lack of universal understanding for many of these terms, they shall be defined for use within this work to ensure common understanding.

1.1.1 Immersion as defined in other literature

The Cambridge dictionary defines the term immersion, when referring to media, theatre and film, as describing the fact of '[becoming] completely involved in something' [16]. The Oxford Dictionary clarifies the term as '(of a computer display or system) generating a three-dimensional image which appears to surround the user' [17]. These definitions may be suitable for the average consumer, however when it comes to research, it does not encapsulate the full depth and complexity of human perception and the whole sensory experience.

Within research literature more broadly, despite the lack of a standard definition for immersion, it is generally agreed that immersion is a multidimensional construct [14]. Various descriptive words have been proposed to communicate different dimensions of immersion, from narrative and ludic [18], to perceptual and psychological [19], to sensory, physical and systematic [20]. These varying terms across different bodies of research largely overlap in meaning, and generally describe the concepts of either 'presence' or 'involvement', as is depicted in Figure 1.1 [21] [22] [23] [20] [24] [18]. This involvement can be passive, wherein a user is immersed through character and story, or it can be active, with the user being physically or mentally engaged with a task.

There have also been many attempts to provide a one-dimensional definition of immersion, however, as these definitions tend to focus on one facet of the immersive experience, such as cognitive aspects, or envelopment within an environment, they did not feel appropriate for this body of work [13] [25]. Outside of the more psychological definitions regarding immersion as perceptual and cognitive, it has also been defined as a technological process. The level of immersion is equated to the level of the technology, with better immersion coming from more advanced technology [14] [26] [27]. However again, this does not really fully encapsulate the idea of immersion within this work.

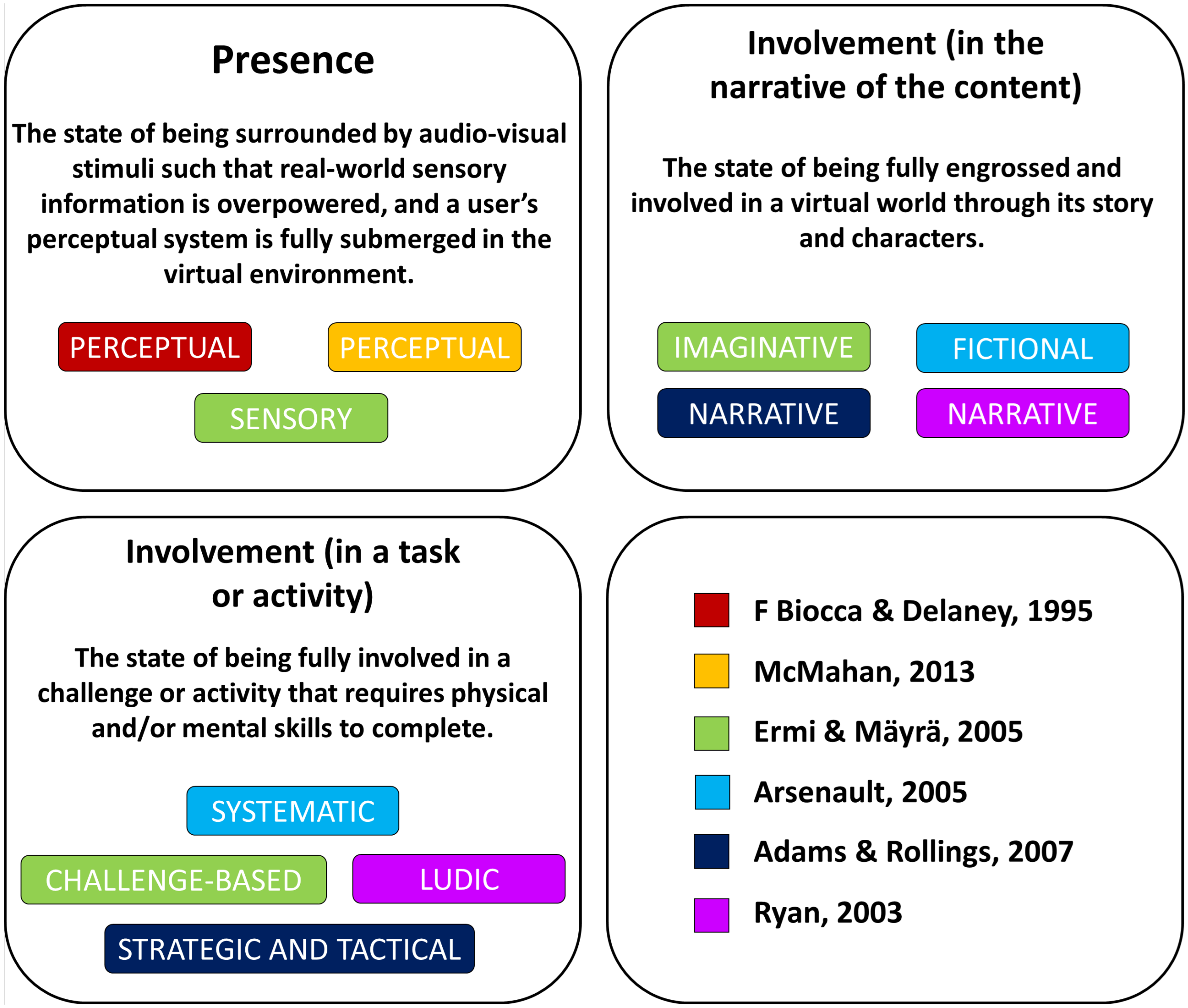

Presence
The state of being surrounded by audio-visual stimuli such that real-world sensory information is overpowered, and a user's perceptual system is fully submerged in the virtual environment.
PERCEPTUAL
PERCEPTUAL
SENSORY
Involvement (in the narrative of the content)
The state of being fully engrossed and involved in a virtual world through its story and characters.
IMAGINATIVE
FICTIONAL
NARRATIVE
NARRATIVE
Involvement (in a task or activity)
The state of being fully involved in a challenge or activity that requires physical and/or mental skills to complete.
SYSTEMATIC
CHALLENGE-BASED
LUDIC
STRATEGIC AND TACTICAL
F Biocca & Delaney, 1995
McMahan, 2013
Ermi & Mäyrä, 2005
Arsenault, 2005
Adams & Rollings, 2007
Ryan, 2003

Figure 1.1: Summary of different immersion terms from the literature.

1.1.2 Definition of immersion

To avoid confusion about what is meant by immersion, it is proposed that for this thesis, set definitions be made to clearly define what is being explored by this work. "Immersion", and "immersive experiences" will refer to the feeling of being fully absorbed in an extended reality (XR) environment emotionally through involvement and presence. When it comes to the technological and physical creation of an immersive experience, this will be referred to as an “immersive system”.

Helping a person to achieve an appropriate feeling of “presence” in a virtual space requires a blend of narrative and environment, alongside physical and sonic immersion. To achieve all, a convincing world needs to be combined with a collaboration of immersive software and hardware. From the hardware side, most VR content is delivered in stereo; however, these two channels can also contain embedded binaural and Head Related Transfer Function (HRTF) tonal colouration relative to where the user is looking and standing in a virtual environment [28].

Stereo-based hardware is not always appropriate however. Headphones are generally designed for listening to entertainment in places where sound may need to be isolated and background noises blocked out, as opposed to the design being prioritised for sonic immersion. In-ear buds or on-ear headphones require contact or surrounding of the ear to achieve optimal sound quality, which can sometimes work against immersion. For example, headphones can create painful pressure over time, pulling a person out of an immersive VR experience; in-ear buds deliver sound directly into the ear canal, which causes listeners to miss out on the effects of their own ears and head interacting with sound waves, which can cause sounds to feel as though they come from inside a person’s head, however spatial they were to start with [29, 30].

There have been developments on the software side through the creation of plug-ins such as Google’s Resonance Audio, which allows developers to have improved sonic positioning power, and to create more accurate virtual reverberation due to features such as virtual sound occlusion, and propagation and reflection [31]. Also, progress is being made in virtual reality hardware to improve aural immersivity. The Valve Index team created a pair of ultra near-field, full range, off-ear (extra-aural) headphones to create a sound field at a person’s ear, and other headset developers are now following suit as it has been shown to increase immersion [28].

Projects outside the world of virtual reality gaming development have utilised various loudspeaker arrays to attempt to improve aural immersion to great effect in recent years [6], [32], [33]. However, there is still room for combining the current and up-and-coming software with hardware from both the virtual reality gaming scene and research world.

1.2 Realism, naturalness and authenticity

The discussion surrounding immersion and immersive systems within this study leads us to an examination of how one might explore the concept of authenticity when recreating or reimagining a real-world space virtually. Later chapters will examine multiple experiments where a real-world space is recreated visually and aurally to try to either mimic said real environment as truthfully as possible, or to recreate the environment in such a way that it still feels like the real environment even if it is not an exact replica. Therefore, a clarification of what is meant by “realistic”, “natural”, and “authentic” is required.

For the purpose of this thesis, “realistic” means as true to the live or real experience and/or environment as possible, trying to fully imitate a real-world aural environment. Whereas an experience being “authentic” is one that is *perceived* as real, one that a user might consider plausible for the real-world equivalent, even if it is not. The word “natural” is also used in this thesis to identify experiences that are real feeling even if they are not entirely real-world accurate. Therefore an “authentic” experience can have inauthentic components or qualities yet still feel like it is true to the real world. A wider discussion surrounding the paradox of inauthenticity being used to improve the “realness” of an experience can be found in Stevens and Raybould’s paper about immersion in a video game series, titled *The Reality Paradox: Authenticity, fidelity and the real in Battlefield 4* [34].

1.3 Thesis aims

The aim of the work presented in this thesis is to explore various methods of auralisation for virtual spaces to investigate not only our ability to recreate virtually the acoustics of real environments naturally in initially simple ways, but to also discover which audio elements are key in improving immersion within these virtual models. The purpose is to understand what prevents a person from feeling aurally, and also generally, immersed in a virtual experience through natural real-time spatialisation and acoustic modelling, as improving the authenticity of a virtual environment leads to greater immersion [35,36]. This research could then inform both researchers and game creators on immersive acoustical aspects to consider when creating projects for virtual reality to guarantee a greater level of immersion from users.

1.4 Hypothesis

The level of immersion experienced by a user in virtual reality, both in terms of realism and naturalness, can be improved by implementing acoustics that are directly based on the real-world environments they are trying to represent, however, the expectations of the user also have an effect on what makes an environment sound real to them.

To test this hypothesis:

- Real and falsified acoustics for an environment will be compared through perception to see if simple audio recreation techniques are suitable enough to simulate real-world recorded audio

- Speakers and singers will be recorded in an anechoic environment and their audio will be implemented into a virtual environment for comparison to speaking in the real space
- An exploration into how changes in reverberant tail and Initial Time Delay Gap (ITDG) contribute to a person's perception of space and sound will be undertaken
- A new device will be designed and built to allow for the real-time manipulation of a live input from a performer through convolution with multiple impulse responses
- Virtual spaces will be rendered to allow for performers to hear themselves singing/ speaking in real time as though in the real performance location.

Three experiments have been conducted to evaluate different aspects of the immersion experience in virtual reality for real-world spaces. The first experiment aimed to explore the effectiveness of simple audio recreation and whether real-world audio experiences could be transparently recreated in VR. The second study explored whether a computer-simulated auralisation for a space was better, equivalent, or worse to a simulation created from at-location measurements. In this experiment, participants could explore the spaces physically in virtual reality. The third experiment explored which sound cues played the most importance when mimicking a real space, allowing participants to play with parameters such as first reflection delay and length of reverberant tail.

1.5 Novelty of research

As will be discussed in Chapter 2, there has been a lot of research into creating immersive spaces for performance in VR, and research into 3D audio for virtual audio spaces. But there is a gap that can be filled which would bring the two together. This is where much of the novelty of this research lies: bringing together experimental exploration of virtual spaces while engaging with adaptable acoustic models and audio sources in the virtual domain, as well as being able to hear oneself as if you were in an acoustically different environment. This is supported by taking a thorough look at what elements contribute most to aural immersivity in virtual environments, and how accurately these need to be recreated for a person to accept them as being authentic. Whether any juxtaposition between visual components and what a person is hearing affects their immersion, even in the aural components are true to a real-world environment. Finally, the creation of a new device which can convolve live audio with multiple impulse responses at an ultra-low latency will be pitched and created.

The major contribution of this research is in defining and quantitatively comparing the set of audio-based factors which create immersive virtual environments. This work is supported by multiple demonstrations, experimental paradigms, and listening tests.

1.6 Structure of thesis

The thesis is organised as follows:

Chapter 2 - Theory and literature introduces core concepts surrounding VR and immersive audio. Section 2.2 discusses how real environments can be recreated virtually, focussing on modelling techniques. The development of sound systems for immersive audio is explored next, leading into auralisation and acoustic modelling. Section 2.6 details the different methods of convolution of impulse responses, followed by a discussion on the recreation of acoustics virtually, focussing once again on the modelling process. Finally, a study of research in the area of immersive audio, and how it has been utilised is presented, highlighting where potential new research novelties may lie.

Chapter 3 - Effectiveness of simple audio recreation describes a study which tests the capability of recreating real-world spaces visually and aurally, and whether these recreated spaces are understood as a “natural” representation of their accompanying real-world environment. The chapter starts with a brief discussion of the hypothesis for the study, followed by the methodology behind creating it. Section 3.3 describes the audio and video rendering processes, then the set-up behind the study and what changes had to be made due to world issues. The results of a questionnaire filled in by participants are stated and analysed in Sections 3.5 and 3.6, as well as further exploration and improvements that could be made.

Chapter 4 - Measuring and modelling physical spaces virtually presents an experiment in which the steps to design and implement a full-sized virtual model of a chapel that allows for real-time changes between acoustic models are described. The goal of this chapter is to compare the perceived validity of auralisations fabricated with acoustic modelling techniques with sound produced by real-world measurements for visually varying virtual spaces. This chapter begins with a brief discussion of the motivation behind this work, then a discussion of the creation of the model itself, both physically and aurally, in sections 4.2 and 4.3 respectively. The method behind the study procedures is outlined, followed by a discussion of the results. Section 4.6 displays a brief overview of the research conducted, alongside potential further research concepts highlighted by the study.

Chapter 5 - The design and implementation of a real-time audio convolution system describes the creation of a novel, real-time audio convolution system which will allow for the real-time manipulation of a live audio input from a performer, granting them the ability to alter the virtual performance space in which their voice is being simulated in. The system created allows the user to change acoustic variables of an impulse response as it is convolved with their own voice through a portable, lightweight device named “The Bela Box”. This chapter begins with an exploration of the use of Pure Data for creating a novel real-time convolution patch for varying impulse responses and audio inputs, followed by a section on the use of Bela as a tool for managing this low-latency audio manipulation. Section 5.4 explores audio processing, looking more specifically at how real-time audio programming can be implemented as well as the limitations of these systems. This section also details how through a mix of frequency domain and direct form convolutions, C++ and Bela can be combined to create a low-latency convolver. Section 5.5 looks at future developments for The Bela Box, followed by a summary of the chapter.

Chapter 6 - Some sound cues contribute more to a person's perception of a space than others details a project which tests how relevant specific sound cues in a reverberant space are for a person to feel as though the acoustics for a space are correct when stimulated by a live input. The chapter begins with an exploration into the topic with a pre-experimnent, which is followed by an explanation for the motivation of the main experiment in Section 6.3.1. The recording processes for capturing the visuals and audio for the choir are detailed, followed by sections describing the rendering and installation required for the study from Section 6.6. The results of the study are recorded and analysed, followed by a discussion of the data, and potential future changes to the system. Finally, the chapter concludes with a summary of the project and the results.

Chapter 7 - Analysis and conclusions brings together the results of all the research undertaken throughout the thesis and examines the evidence that supports the main hypothesis. Section 7.3 includes suggestions for further work that could be undertaken within the sphere of immersive audio and virtual reality, detailing specific areas in which this could be applied. A summary of potential areas of application that this work is relevant to is summarised in Section 7.4, and then this chapter concludes with some final concepts regarding the thesis and discoveries as a whole.

Chapter 2

Theory and Literature

“Why shouldn’t people be able to teleport wherever they want?”

- Palmer Luckey, *Founder of Oculus VR* [37]

In order to create physical spaces in virtual reality authentically both visually and audially, it is first necessary to gain an understanding of what virtual reality is, fundamental properties of sound propagation, the different ways in which audio can be presented to a listener, as well as current projects exploring facets of this area of research.

This chapter will open with a brief introduction to virtual reality and methods to recreate physical environments virtually. Section 2.3 details the development of sound systems, both generally and for VR. Sections 2.4 to 2.6 explore the auralisation of real environments, how we can model sound for VR, and how convolution of Room Impulse Responses (RIRs) can be implemented to recreate acoustics of spaces. This chapter finishes with a discussion of previous immersive audio experiences that utilise the highlighted techniques to explore VR and immersive audio within research.

2.1 Virtual Reality

Virtual reality (VR) technology, and our access to it, has come a long way over the last 30 years. Prior to the 1990s, the early versions of VR devices as we know them today were researched and produced mainly for military and medical purposes [38]. From then onwards, the first widespread commercial releases of consumer headsets began, with multiple gaming companies such as Sega and Nintendo turning their focus towards this new gaming avenue [39, 40]. However, it was not until the mid-2010s that VR really took off with the general public, spearheaded by Oculus' Kickstarter campaign [41].

VR can create a simulated environment within a three-dimensional experience that is able to convince the user that they are in some other place that is not where they are physically. Whether that be a recreation of a real place or a completely fictional one. The simplest of these experiences use specially designed rooms covered with multiple large screens, however the more full-body immersive systems use virtual reality headsets, allowing a user to travel wherever they want. These virtual environments strive to accomplish realistic and immersive experiences through two main aspects: intuitive interactions, and immersion.

Due to the popularity of VR, the former of these aspects is being achieved readily, with the major VR companies constantly competing to improve their technologies. Modern day VR revolves around the use of a Head Mounted Display (HMD) and one or two controllers, all of which movements can be tracked, most commonly through a laser-based inside-out tracking system. A pair of Base Stations or Beacons flood the room with infrared lasers, one axis at a time (left to right, top to bottom), preceded by an emission of a powerful flash of IR. The difference in time between the flash hitting the tracked devices and the scans determines your position in the space [42]. Unlike Valve and HTC, Oculus prioritise the creation of standalone headsets that do not require external tracking devices to track the headset and controllers' locations.

The ocular technology for HMDs are comprised of a stereoscopic display and a pair of lenses to adapt that display for the user [43]. The higher the refresh rate and resolution and wider the field of view; the clearer and more authentic the experience is. Controllers also help with the intuitiveness of interaction. Not only are the positions of your hands tracked, newer controllers even have separate finger tracking, as well as grip sensors, allowing you to pick up objects virtually by physically grabbing and not just by pulling a trigger, thus creating a more 'natural' experience [44].

As discussed in Chapter 1, immersion is hard to quantify as it can mean different things to different users. For some, having a game or experience being as true to the real world as possible creates that immersion. The most recent big VR game release relies on real-world physics in the games to recreate real-world interactions more authentically in VR: heavy things are hard to pick up and move; you have to pull a door open, not just click on it for example [45]. Furthermore, sounds you make in the real world can be integrated into game mechanics. Breath Tech is a VR puzzle game that uses your breath as an integral mechanic. You can blow out candles or make bubbles underwater through the inbuilt mic in the headset. The creator Brett Jackson even talked about immersion in a developer blog, stating,

> "*The more that we represent a user's physicality in the virtual world the more natural it feels. You've experienced it with your hands or seeing your shadow/reflection move with you, it's the same with your breath*" [46].

Sound also plays an integral part in immersion. If you see that you are in a forest for example, you expect to hear wildlife, the leaves, the ground crunching underfoot. If you see a car pass in front of you from left to right, you expect the sound of that car to also travel from your left side to your right. In *Half-life: Alyx*, there is an enemy that is blind, but can hunt its prey due to acute hearing. If you knock something over in-game, it will make a sound, and the beast will find you. There are even spores in the air in some areas that will make you cough if you do not cover your mouth physically with a controller, causing the beast to find you [45]. Immersive, reactive audio and its development in both VR and generally will be discussed in further detail within sections 2.2 to 2.5.

Although not integrated as of yet into household VR experiences, other areas of media have attempted to account for other senses, such as smell and full body sensations like temperature within virtual experiences. 4D cinemas and rides are becoming more and more popular for example. If there is an earthquake, the seats may rumble. If a monster sneezes on you, water is sprayed towards you. If a scene pans to the sea, you may smell the ocean [47]. As can be seen, virtual reality encompasses many areas which strive to make a person feel more immersed in a situation. For the future of this work, a focus on computer-driven headset-based virtual reality for the visual side of immersivity will be made.

2.1.1 Game engines

When it comes to programming in virtual reality, there are two main competitors: Unity and Unreal Engine. Both are cross-platform game engines which allow the user to create three-dimensional, two-dimensional, and virtual and augmented reality experiences for games, or for more industrial areas such as film, engineering, and architecture.

Both engines have their differences and benefits despite their similarities:

Benefits of Unreal Engine

- Directly implements C++
- Utilises 'Blueprints': a visual coding system for ease of programming
- Includes version control
- Has a more powerful and sophisticated rendering system, i.e. smoother lighting and shadows
- Has a material editor for tweaking and creating materials
- Has multiple plugins for creating immersive audio in VR that take into account direct sound and reflections

Benefits of Unity

- Can implement classes using C#
- Has extensive documentation for using the programme
- Has a huge assets store for creating rooms in VR
- Far easier to learn due to vast content online and intuitive interface
- Has built-in basic ambisonic audio implementation
- Has multiple plugins for creating immersive audio in VR that take into account direct sound and reflections

Although there are multiple ways to approach the creation of physical environments virtually outside of a traditional game engine, such as 360° video players, or using programming languages such as Java for phone-based VR experiences; for PC-based VR, Unreal Engine and Unity are the most useful. When it comes to the creation of visual virtual spaces, there are multiple approaches that can be taken that are applicable to both engines.

2.2 Virtual recreation of real environments

There are multiple tools through which we can visually model physical environments virtually.

2.2.1 3D modelling

One method is to physically model the environments that are being recreated. Programs such as Blender, Inventor, and general CAD software can be utilised to build 3D models that can be exported into virtual reality game engines. Unity, one of these game engines, has in-built experimental building tools that can be used to create simple 3D models in engine.

2.2.2 Camera-based modelling

3D modelling can be time consuming, although precise. Another method through which a space can be virtually modelled is via camera-based techniques, where spherical cameras are used to reproduce real spaces in virtual reality. A pair of 360° images are taken to create a simplified 3D geometric model of the scene, estimated through depth estimation from the captured images and semantic labelling using a convolutional neural network (CNN) [48].

2.2.3 LiDAR

LiDAR (Light Detection and Ranging) scanning technologies can be used to quite quickly create accurate 3D models for virtual reality which can also be combined with 2D imagery to construct photo realistic models of 3D spaces [49–52]. Newer commercial phone and tablet products have in-built LiDAR technologies, so a physical space can be scanned a single time with this range scanner and when paired with videos and photos of the scene, a 3D representation of the physical location can be created [53].

2.2.4 360° Imagery and video

For some situations, a fully rendered space to move about in is not required. For a stationary experience where the user can only move on the spot, 360° cameras can be used to capture real locations to be viewed.

It is important to remember that when a photo is stretched over a sphere, it effectively loses resolution. A regular 4K photo for example will result in a clear, high resolution image, but over a sphere, it will appear to be at a lower resolution. There are two main options for camera setups for 360° imaging. Either a cube of cameras is required pointing outwards to record 6 directions at once, such as the GoPro Omni Kit, or a single camera that can record two 180° images at a time, such as the GoPro Fusion is required [54]. As 180° camera resolution is improving, either type is suitable although a higher final quality and resolution is generally achieved through the cube set-up as each image is less stretched when mapped to a sphere.

These spherical images can then be handled in a variety of ways. Both Unity and Unreal Engine use a skysphere, where the imagery is mapped onto a sphere that is infinitely far away from the spectator. Other programs can just play 360° videos natively if processed and exported in the correct format. Vive Cinema, which will be discussed later in more detail (see Section 2.3.4 and Chapter 3), can play 360° scenes around the observer [55].

This section has focused on how we can recreate physical environments visually, ranging from 3D modelling to photo-realistic methods. However, visual effects only contribute part of what is required to fully model a physical space digitally.

2.3 VR and immersive audio - development of sound systems

There has been extensive research into what makes virtual spaces more immersive [34,56–62], and most conclude that audio plays a vital role in not only making experiences *more* immersive overall, but also more realistic. Multiple of these sources reference how the addition of sound in a simulation helps add to the illusion that the depicted scenario is actually occurring by transforming their sense of place. Eaton and Lee go on to explore which audio factors contribute more to a person's sense of immersion more specifically, demonstrating that, for example, horizontal sound perception was perceived to be more important than vertical perception of sound [56]. This section explores the development of sound systems for virtual reality, describing speaker systems and their personalisation to improve immersion.

2.3.1 Head-Related Transfer Functions (HRTFs)

Spectral cues, interaural level differences (ILDs), and interaural time differences (ITDs) are the main ways in which humans localise sounds in a space, and a Head-Related Transfer Function contains all of this information [63]. As a sound wave reaches a listener, their pinna (outer ear) and ear canal, as well as their body and head, can all alter that sound's acoustical properties before reaching the listener's eardrums [64, 65]. These modifications are typically discerned as adjustments to a source's position in relation to the listener's head such as elevation and front/rear localisation, even though HTRFs also modify the spectrum of the original source. The function which describes this acoustic transfer function between a point source in the free-field and the listener's ear canal is called the Head Related Transfer Function (HRTF) [66].

HRTFs are unique to each person, being based on their specific anatomy, and are therefore time-consuming to measure. Multiple methods have been established over the years to measure and model HRTFs. For an in-depth overview of these, a reference has been provided [67]. This thesis did not measure or create any HRTFs as part of this research, however the function has been highlighted as it can play a key part in improving personal immersivity in virtual aural environments.

Binelli's team were the first to utilise a modified version of Vive Cinema: an open source project compatible with many HMD systems, with a friendly user interface [55]. Vive Cinema can play high resolution 360° videos, and can manage spatial audio formats up to 3rd order Ambix. It also uses generic HRTFs, which cannot grant to everyone the same localisation capability as this is affected by head dimensions and shape: the qualities that are generalised in this system. The team wanted to create a ready-to-use application for comparing HRTF sets through 360° audio and video. As the software is open-source, it was easy to modify, allowing for the group to edit features of the program, removing the in-built symmetricity hypothesis, replacing the mid/side convolution scheme with their own full matrix one. Vive Cinema uses this mid/side scheme to reduce computational load for binaural rendering, but it relies on the assumption that HRTFs are perfectly symmetrical. The whole modified project (source code and precompiled executable) is available for download at the following link: http://www.angelofarina.it/Public/ViveCinema [55].

2.3.2 Spatial audio system development

As this thesis is investigating the use of VR and spatial audio techniques to simulate real-world acoustic environments, the following chapters will focus mostly on the use of ambisonic and binaural audio as they enable the delivery of three-dimensional sound over headphones. However, the development of surround sound and spatial audio systems are a useful starting point for further elaboration.

Mono

The original available speaker systems were monaural; using one microphone to record a sound scene, then using a single speaker to reproduce all of a recorded sound without any separation [68]. One signal, one channel. Even if played over multiple speakers, the audio output is the same from each one. It is stated by Steinberg and Snow that a single channel *can* preserve some of the depth qualities of the recorded scene, but it is preserved imperfectly and does not fully capture the essence of the original sound field [69].

Stereo

One of the simplest and most common systems used to simulate spatial audio is stereophonic sound. This is achieved by utilising two or more audio channels and playing them through a configuration of two or more speakers (or a pair of stereo headphones). Two signals, two channels.

Today's practical stereo systems are derived from the work of Alan Blumlein in the 1930s, and have been in use for decades, meaning that they have been extensively studied [70]. In general, stereo sound relies on our ability to position a sound between two speakers, which allows for sound sources to appear to be heard from various directions, like regular hearing. This is done by adjusting the delay of the sound (time panning), and/or the amplitude (intensity panning) at each speaker [71]. These techniques can even create the impression of virtual 'phantom sources' that do not appear to come from the speakers [72]. This implementation is limited however as, although it can simulate spatial qualities, it does not provide height cues.

5.1/7.1 Surround

Following stereo came quadraphonic sound systems; 4 speakers in a square arrangement, two in front of the listener and two behind, which was introduced to the public in the early 1970s [73]. This was created in an attempt to have a more immersive audio experience and was the earliest consumer surround sound product. However, it still had its issues. It is difficult to reproduce directions between the speakers, especially the lateral ones. Furthermore, broadcasting and placing four tracks on two track stereo media also caused problems, and although matrix encoding of channels did work, when trying to create sounds that come from the sides of the listener, there appeared to be a frontal dominance induced [74, 75].

From spatiality though, discrete 5.1 surround sound systems were created. This setup is comprised of three frontal channels, one at 0° and two at ±30°, two surround channels at ±120°and a subwoofer. Figure 2.1 demonstrates the layout.

The central speaker provides an anchor for the listener, meaning that even off-centre listeners can still feel the spatialisation effects created from whatever media is being listened to. 7.1 surround systems add two extra speakers to the set-up at ±90° from the listener. This is done to stabilise panning effects to the side of the listener, as in 5.1, the gap between the front and back speakers is large, causing sounds to appear to jump between the front and rear channels [72].

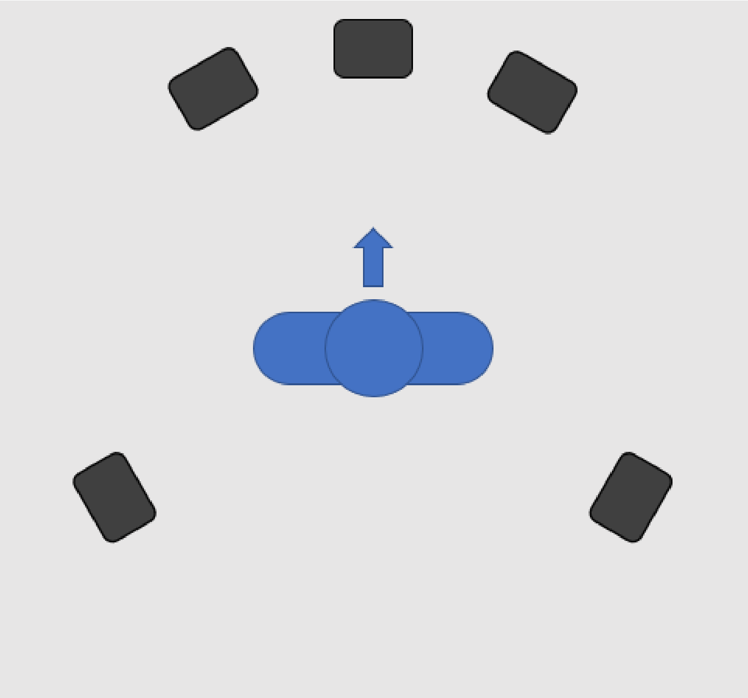


Figure 2.1: A diagram of a 5.1 surround sound system layout, where the arrow indicates 0 degrees. The blue figure in the centre of the image represents a listener, and the five grey boxes the speakers.

2.3.3 Binaural

Although surround sound and even stereo can be used effectively to improve how immersive a non VR experience feels [76], neither of these methods account for height of sound. Virtual reality brings different challenges to immersing a user aurally as the listener can move within a space, even moving below or above potential audio sources. It has been shown over multiple studies that spatial audio fidelity has a particularly significant effect on immersion in VR.

For example, O-larnnithipong et al. explored the use of binaural 3D sound compared to sound whose intensity (equal in both ears) decreased as the straight-line distance from a sound source increased, for finding a sound source in Virtual reality [1]. Two symmetric mazes were created in Unity to test the two audios, referred to as “3D SOUND” and “2D

SOUND" respectively (as for the latter auditory condition, the amplitude does not depend on the navigator's head orientation, only the x and y coordinates of the person). These mazes can be seen in Figure 2.2.

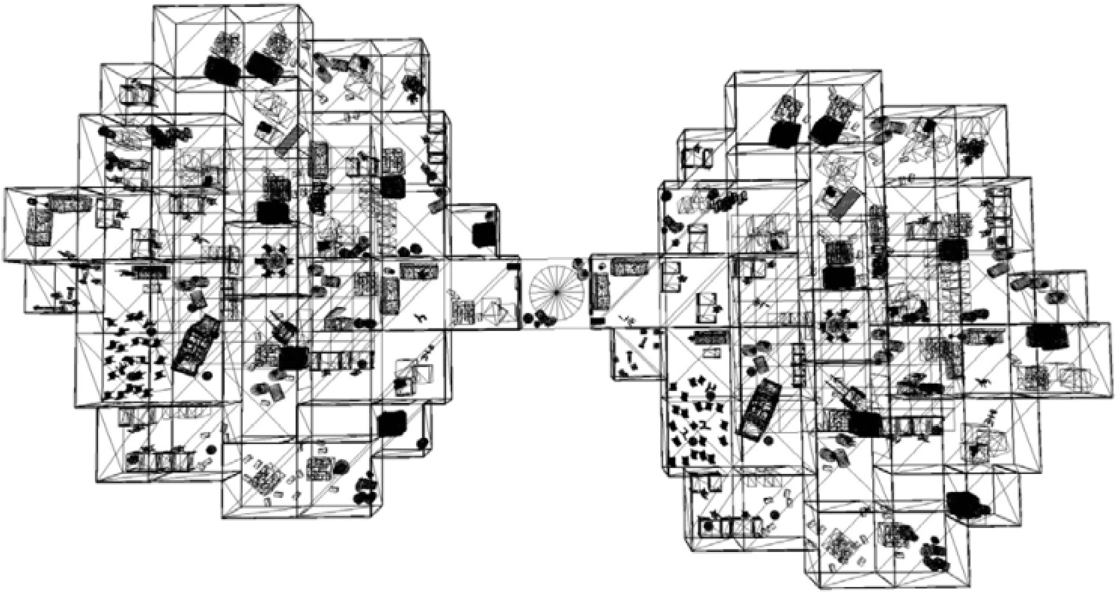


Figure 2.2: Symmetrical mazes for testing the two audio conditions. For each test, the participant would start in the centre [1].

The 8 participants found the sound source far quicker with the fully binaural sound type, with the meantime being 244.18 seconds faster than the 2D sound maze. These results provide strong confirmation that head tracking with directional audio plays a very important role in realism of contemporary immersive VR environments.

Furthermore, a study performed by Potter et al. in 2022 demonstrated that spatial audio was so important for improving immersion in VR that "the addition of room acoustic rendering to head-tracked binaural audio had the same improvement on immersion as increasing the video resolution five-fold, from 0.5 megapixels per eye to 2.5 megapixels per eye" [77]. Subjects wearing head-mounted displays and headphones were presented with a virtual environment with music and spoken audio sources using three tiers each of video resolution and spatial audio quality. The audio was rendered monaurally, binaurally with head-tracking, and binaurally with head-tracking and room acoustic rendering. The results showed that both improvements to video resolution and to the audio immersive system had a significant effect on user immersion, with audio playing a larger role in this improvement of immersion.

2.3.4 Ambisonics

Ambisonics is a method of spatial audio reproduction that was first introduced in the 1970s that can be used with speaker arrays, but also through headphones binaurally. It provides a full sphere of surround audio, including height ("periphony") while only requiring 4 total channels [78], and is based on spatial sampling and reconstruction of soundfields using spherical harmonics [79].

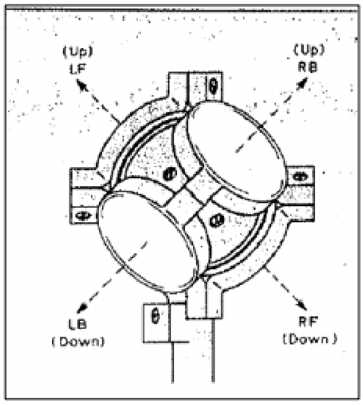
Encoding and decoding of ambisonic audio will be touched upon in this section, with more in-depth resources provided. This is because the encoding and decoding of recorded audio throughout this thesis was handled by the Zoom H3-VR first-order ambisonic recorder itself [80].

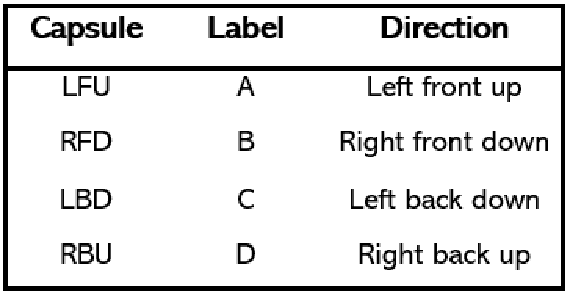
Unlike the other multichannel surround formats mentioned above, Ambisonics can be decoded to any speaker array as opposed to a pre-determined array, making it incredibly versatile. This is because through encoding into Ambisonic format, the soundfield is broken down into orthogonal functions, and the weighted combination of channels can produce sound in any direction. This also means that we can very easily rotate or transform the soundfield, which makes it useful for VR experiences [81, 82].

First-order ambisonics uses 4 channels. There are two kinds of first-order ambisonics; A and B format. The 4 signals outputted from the microphone as shown in Figure 2.3 are known as the A format. These can be converted to B format by using a combination of the signals. This conversion is achieved by,

$$\begin{aligned} W &= A + B + C + D \\ X &= A + B - C - D \\ Y &= A - B + C - D \\ Z &= A - B - C + D \end{aligned} \tag{2.1}$$

where the W channel is omnidirectional, and the X, Y, and Z channels are Cartesian direction figure-of-eight polar patterned.





Capsule	Label	Direction
LFU	A	Left front up
RFD	B	Right front down
LBD	C	Left back down
RBU	D	Right back up

Figure 2.3: Soundfield Ambisonic Microphone capsules [2] and the names of each capsule.

A monophonic sound signal can be encoded into B format first-order ambisonics by taking the source signal and azimuth (θ) and elevation (ϕ). This signal is then distributed over the ambisonic components with the B format gains for each channel,

$$\begin{aligned} W &= \frac{1}{\sqrt{2}} \\ X &= \cos\theta\cos\phi \\ Y &= \sin\theta\cos\phi \\ Z &= \sin\phi \end{aligned} \tag{2.2}$$

Decoding a B-Format Ambisonic audio file for a loudspeaker array requires a decoding matrix [81], [83]. For an arbitrary number of regular, equally spaced loudspeakers in a speaker array, a simple sampling decoder [84] can be used to calculate the gain for each Ambisonic channel based on each loudspeaker's position, denoted by θ_l and ϕ_l;

$$\begin{aligned} g_W &= \frac{1}{\sqrt{2}} \\ g_X &= \cos\theta_l\cos\phi_l \\ g_Y &= \sin\theta_l\cos\phi_l \\ g_Z &= \sin\phi_l \end{aligned} \tag{2.3}$$

with the resulting loudspeaker signal s_l [85] being given as

$$s_l = \frac{(2-d)g_W W + d(g_X X + g_Y Y + g_Z Z)}{2}, \tag{2.4}$$

where d is the directivity factor of the virtual microphone response, in the range 0 ≤ d ≤ 2, such that d = 0 results in an omnidirectional virtual polar pattern, d = 1 results in a cardioid polar pattern, and d = 2 results in a figure-of-eight polar pattern. At the centre of a regular, equally spaced speaker array, for low frequencies, the original soundfield can be accurately represented [84]. This thesis predominately focuses on first-order ambisonics.

2.4 Auralisation of real environments

The standard definition of sound is that it is an oscillation in pressure, stress, particle displacement, particle velocity etc., propagated in a medium with internal forces or the superposition of such propagated oscillation [86]. Humans are sensitive to variations in pressure which occur within the frequency range of 20 Hz to 20.000 Hz. In an open space with nothing to interact with, sound waves propagate outwards in three dimensions, and the sound intensity decreases with distance, following the Inverse Square Law

$$I \propto \frac{W_{source}}{4\pi r^2}, \tag{2.5}$$

where I is intensity (W/m^2), W_{source} is the power of the source (W), and r is the distance from the sound source (m^2).

In a room though, the sound waves that transmit through the air to the listener can reflect and scatter off of surfaces and objects present in the room, transforming the heard sound. Therefore, this sound is heard as a subset of sound waves that have combined. Firstly, we have **direct sound**, which travels straight from a source to the listener, and is the first element to arrive at the ear. Next, some **early reflections** arrive soon after. These sound waves have bounced off a few surfaces before arriving at the listener very shortly after the direct sound. The first order reflection is also lower in amplitude than the direct sound as it has reflected off a surface before arriving at the listener. The next early reflections to arrive will have decreased in sound intensity following Equation 2.5, but the timings and frequencies of these later reflections will vary due to the positioning of walls/ceilings and the reflective qualities of the room [87]. Later, reflections will arrive at the listener that have a far lower amplitude and are temporally closely spaced. These are called the **reverberant tail**. The duration of this late sound is described by the reverberation time, which is defined as the time for the sound level to decay to one millionth of its energy [88].

Generally, a person will judge the position of a sound source based on which direction the direct sound arrives, whereas the balance of direct and early sounds helps a listener judge the distance to the source, so even if we do not consciously consider the early reflections, they are important for discerning locations of sound sources. How close the listener is to the source also affects how the sound will be perceived. If close to the sound, the direct sound will dominate for the listener, and if further away, the early reflections will arrive far closer to the direct sound as both have travelled a similar distance [88].

2.5 Acoustic modelling: Impulse Responses

Many acoustical parameters for a space can be derived from an accurately measured impulse response which can then be integrated into a complete auralisation process. A good signal-to-noise ratio (80 dB or higher) is required if a high quality result is to be gained [89], [90]. An ‘ideal’ impulse is defined to be a signal of infinitely short duration, infinitely high power, and unit energy, however, this Dirac function produces an exact impulse response that is only theoretical [91]. Nevertheless, there are achievable requirements we can satisfy to attain the best impulse response.

2.5.1 Composition of an Impulse Response

The Room Impulse Response (RIR) can be considered the unique “acoustic signature” of a room, representing how the room reacts to a specific excitation signal with a given source and receiver configuration [92]. In an enclosed space, a portion of the emitted sound source will reflect off the room boundaries and gradually diminish as they get absorbed by the surfaces or air in the room. This RIR consists of the components shown in Figure 2.4.

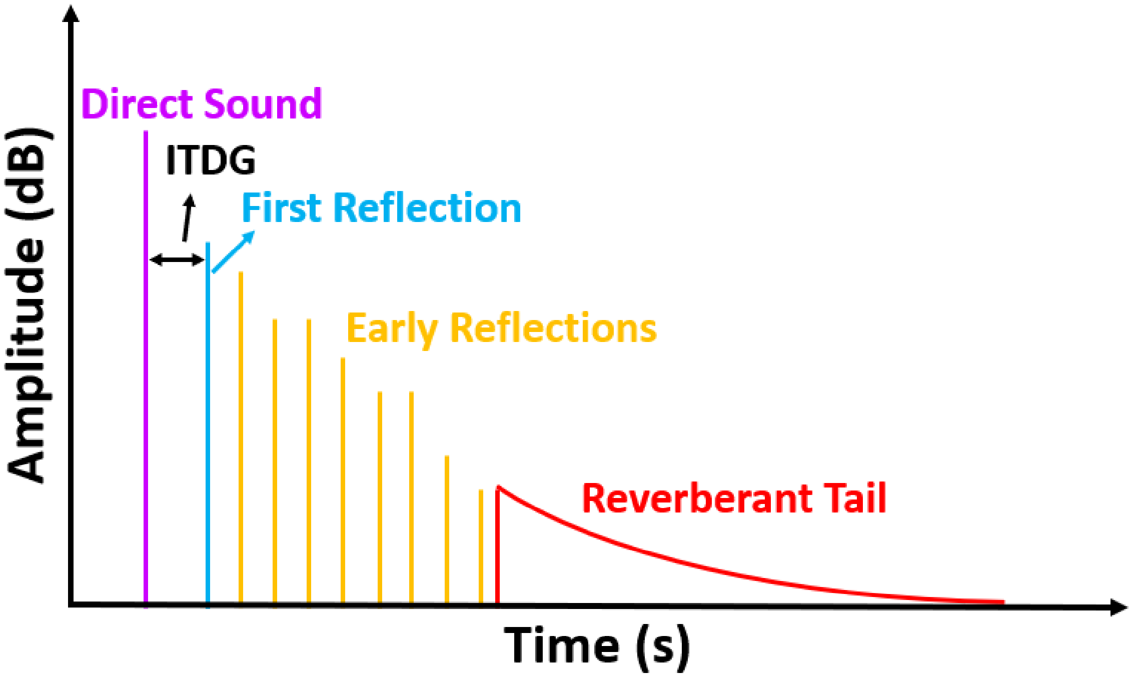


Figure 2.4: Representation of a Room Impulse Response.

Like the descriptions in Section 2.4, an impulse response reacts in a similar way as any other sound source does in a space. Direct sound reaches the listeners' ears directly from the source without reflecting off any surface [93]. This part of the sound has the largest amplitude as it has lost the least energy. The Initial Time Delay Gap (ITDG) is the time delay between the direct sound and the first reflection.

The first reflection and early reflections are the sound waves that have bounced off a few surfaces before reaching the listener, arriving within typically 10 to 80ms of the direct sound [94]. Lastly, we have the reverberant tail, or the late reverberation. This refers to a disordered sound field composed of diffuse reflections [94]. A suitable level of late reverberation can enhance the perception of spatialisation, but excessive amounts can compromise sound clarity.

The best excitation signals should fulfil the following stipulations:

- The emitted source must be reproducible and well-defined
- The signal-to-noise ratio should be maximised
- The source and receiver should be located such that they reflect real positions (i.e., for a concert hall, the source could be located on the stage and the receiver in the audience.)

There are two main methods utilised in the literature when accurately simulating the acoustics of real, physical spaces through excitation signals: using a well-defined signal or making a sound and recording the response at the listening position to find the RIRs.

2.5.2 Room Impulse Response from well defined signals

Impulse responses can be calculated from the response to another known, well-defined signal with sufficient energy at relevant frequencies. There are multiple methods with varying pros and cons that can achieve this. White noise is one of them, however due to the random nature of this excitation signal, the extracted impulse response will show residual noise [95]. This is reducible through increasing measurement time; however it is an inefficient method.

Another example is the Maximum Length Sequence (MLS) signal. This technique employs a periodic pseudo-random signal with almost the same stochastic properties as white noise to excite the acoustical space [96]. Its frequency spectrum over one period is as flat as that of the ‘ideal’ impulse, therefore, there is no noise pollution induced on the extracted impulse response from this excitation signal. Thus, in the presence of non-white noise, the MLS technique is fairly accurate compared to other techniques.

Acoustic impulse responses can also be captured through an exponential swept sine technique [97]. This method allows for measurement of the RIR and any distortion present in the system can be removed. A logarithmic sine sweep is synthesised with constant amplitude which increases in frequency per time unit. This resulting sweep moves fast through high frequencies and slowly through low frequencies.

The sine sweep is output to the room, recorded and then deconvolved with a time-inversed filter (which accounts for the amplitude envelope) of the original input sine sweep. This results in a linear impulse response with an initial delay equal to the length of the input signal. This method is most appropriate in quiet environments. For a more detailed explanation, refer to reference [97]. This technique is used commonly in in-situ room impulse response measurements, for example in churches, and other performance spaces (e.g. [6]). However, this method can sometimes be inappropriate or infeasible for the task at hand due to technology requirements and a different approach is needed. For example, when creating an RIR of a virtual space, as described below.

2.5.3 Room Impulse Response from an excitation signal

Impulsive excitations, such as a gunshot or balloon pop, can be used to record RIRs without any post-processing as the recorded sound in the space represents the RIR directly in the time domain. This method can be utilised for both real and virtual spaces. For example, in *Reproducing Real World Acoustics in Virtual Reality Using Spherical Cameras* [98], comparisons of RIRs for a real and virtual space were needed. The packages used to handle ambisonic audio in Unity faced difficulties when attempting to reproduce sinusoids. Therefore, the sound of an anechoic gunshot (normalised in the time domain) was placed at the source position, and the response recorded at the listening position.

This method of excitation has poorly defined directivity and is difficult to reproduce exactly on repetition. However, as will be shown in Chapter 3, when it comes to people's expectations of the acoustics of a space, this did not seem to make a difference.

2.6 Convolution of Room Impulse Responses

Convolution has many applications in audio and signal processing, including artificial reverberation and equalisation [99, 100]. Convolution of an RIR with a sound source can create the effect of that sound source sounding as though it came from a different acoustic environment. The convolution can be calculated in either the time domain, utilising Finite Impulse Response Filters (FIRs), or in the frequency domain, using Fourier transformations (FFT) [87]. These filters are composed of a series of coefficients whose output is produced as a linear combination of the past and current outputs [101]. The discrete convolution between two signals *x[n]* and *h[n]* is given by

$$y[n] = \sum_{k=0}^{N-1} h[n] \cdot x[n-k], \tag{2.6}$$

where N is the number of samples in $h[n]$, often referred to as a filter. This convolution is commonly rewritten more concisely using $*$ to denote the convolution as

$$y[n] = x[n] * h[n]. \tag{2.7}$$

Room reverberation can be modelled as an FIR filter given a sufficient number of filter coefficients through measuring the impulse response of a room.

2.6.1 Direct form convolution

Once an IR is recorded, it can be convolved with an incoming signal. A common way of implementing an FIR filter is through using direct form convolution

$$y[n] = h_0x[n] + h_1x[n-1] + h_2x[n-2] + h_3x[n-3] + \cdots + h_{N-1}x[n-N-1], \qquad (2.8)$$

where each output of the filtering process $y[n]$ is computed as given in the previous summation. This process is shown graphically in Figure 2.5.

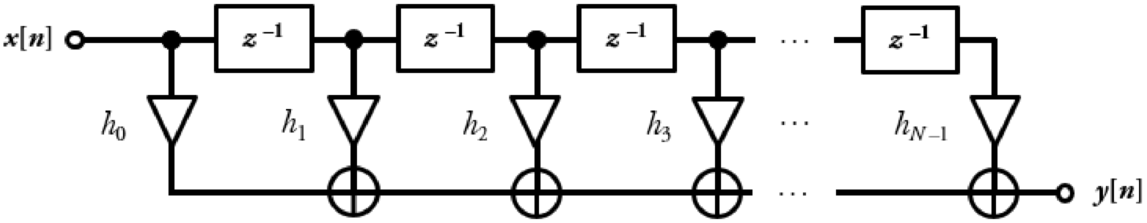


Figure 2.5: A block diagram for the Finite Impulse Response (FIR) filter implemented with a direct form convolution, where *x[n]* is the input, ▷ is amplification, z^{-1} is delay, ⊕ is summation, and *y[n]* the output.

For small numbers of coefficients, this convolution method is quite efficient, however as N increases, the efficiency decreases with a linear scaling of $O(N)$ multiples per output point, as direct form convolutions process over a single sampling period which is equal to one divided by the sample rate. For a sample rate of 44.1 kHz, a sampling period of 22.6 µs is calculated. This value is unrelated to N, so if the number of coefficients is increased, these multiples still need to occur within the sampling period.

2.6.2 Frequency domain convolution

The complexity of this problem can be significantly reduced by taking advantage of the convolution theorem, which states that multiplication in the frequency domain is equivalent to convolution in the time domain [102]. This theorem can be applied here, allowing for the convolution to be computed by initially transforming the input signal and filter to the frequency domain,

$$\mathcal{F}\big[x[n] * h[n]\big] = X[m] \cdot H[m]. \tag{2.9}$$

Therefore, the complex multiplication of $X[m]$ and $H[m]$ will produce the convolution of the original signals in the frequency domain, and by taking the inverse Fourier transform, the time domain output of the convolution can be recovered.

The application of the convolution filter in the frequency domain enables the use of the Fast-Fourier transform (FFT), resulting in a logarithmic scaling of the multiples per output point $O(log(N))$. This results in the process becoming significantly more efficient than the direct form convolution as N increases.

To implement frequency domain convolution, a block convolver is normally used. Input signal $x[n]$ is repeated infinitely on both sides (positive and negative n) every N samples. In order to avoid time aliasing, an FFT size twice the size of the input block is implemented, filling the latter half of the time domain input with zeros, as shown in Figure 2.6 [103].

While frequency domain convolution is more efficient for larger values of N (generally greater than 128 samples) when compared to the direct form convolution, a minimum latency of N samples is imparted, dictated by the FFT block size. Unlike in direct form which only needs a single input sample to produce the next output, N samples need to be accrued from $x[n]$ prior to performing the computation [101, 104].

The issues raised about these two methods of convolution are only apparent when at-

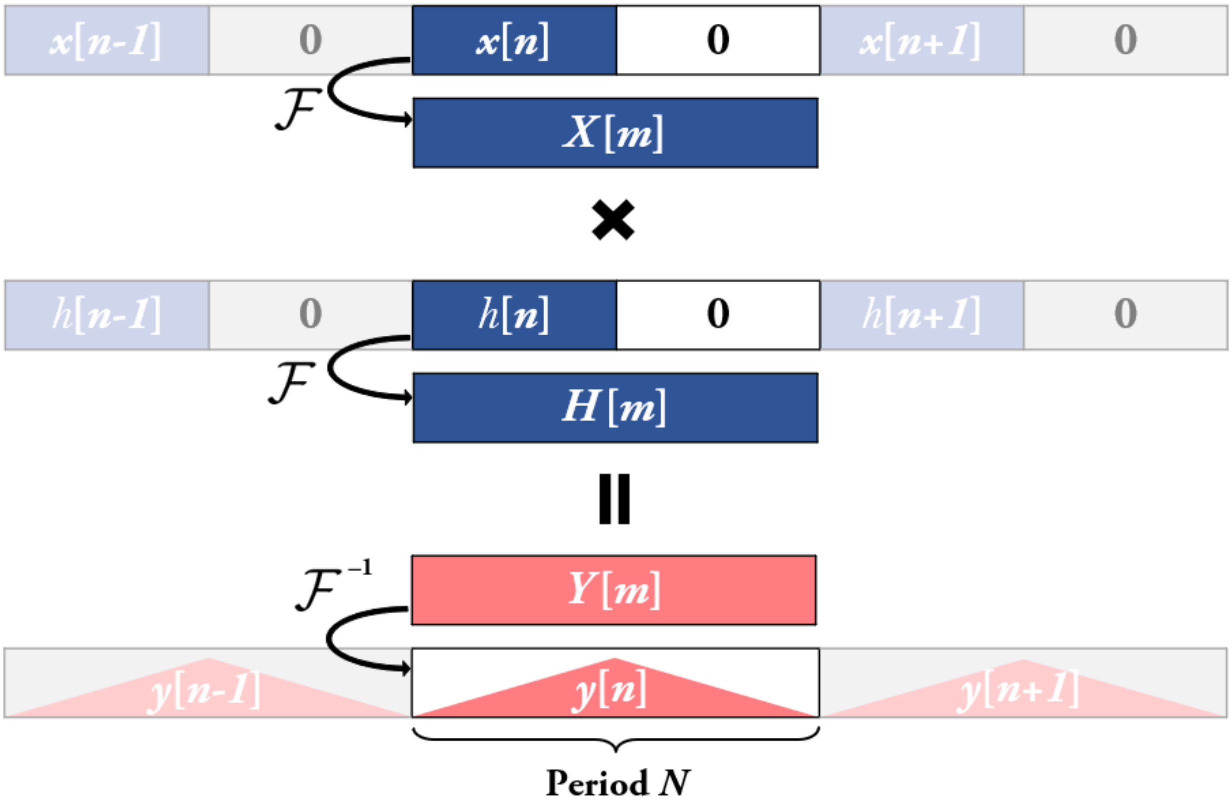


Figure 2.6: A visual representation of frequency domain convolution.

tempting convolution in real time. As will be described in later chapters, there are multiple software systems that can be used to convolve impulse responses with a live audio input at higher latencies, and in Chapter 6, an implementation of both convolutions together is used to create a "zero latency" convolver.

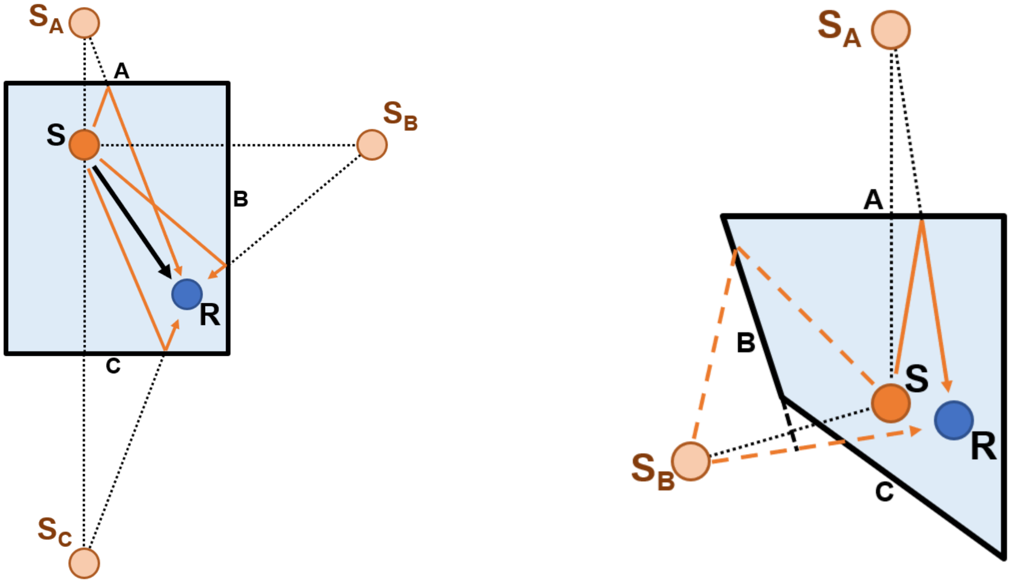
2.7 Acoustic modelling: computational

While IRs recorded in real-world locations can be used to convolve virtual sound sources, the IRs are specific to the exact physical environment as it was in the moment that they were sampled from. For an interactive virtual environment with moving sound sources or objects, it may be too computationally strenuous to sample and load the IR of the virtual space prior to running the program. Through techniques that model sound propagation, the IR of a virtual space from a listener to any sound source can be computationally calculated. These techniques can be separated into two categories; those based on a numerical solution to the underlying wave equation for sound propagation, and those based on sound field decomposition via geometrical acoustics [105–107].

There are multiple numerical and geometrical methods for acoustic modelling that can be, and are, applied when making virtual experiences more captivating. In geometrical methods, a number of simplifications approximating sound propagation and reflection are made. The main ones take advantage of the similarities between light and sound waves, treating them as a ray, or a particle [108,109]. Propagation of these sound sources are then dictated by the geometrical laws of optics, therefore, these techniques are only appropriate for approximating scenarios where the wavelength of the wave is very small in comparison to the dimensions of the space, i.e., a room. [110]. These sound field decomposition techniques calculate the IR as a sum of secondary source contributions to the direct path from the environment. A significant benefit of geometric acoustics is the ability to assign different priority levels to these earlier and later contributions to the impulse response. This can be exploited and employed for calculating high-quality early reflections with only an approximation of the late reverberant field when the simulation's complexity would be far more complex [105].

2.7.1 Geometrical acoustics: image source

The image source method relies on simulating virtual projections of a sound source over all surfaces in an environment to algorithmically find all of the specular reflection paths between that source and a receiver [111]. If a ray from the sound source reflects off of a boundary before reaching the listener, a virtual reflected source, or *image source* is created on a line perpendicular to that boundary and the same distance away as the original source. Figure 2.7a depicts this pictorially for a geometrically simple space, where sources S_A, S_B, and S_C are created when rays from source S collide with walls A, B, and C respectively.



(a) An example of the image source method where all three sources S_A, S_B, and S_C are valid and visible.

(b) An example of the image source method where not all of the sources are valid and visible. Resultant image source S_B is not valid.

Figure 2.7: Examples of the image source method for geometrically simulating a sonic environment.

All reflections are treated as specular, therefore the strength of each new reflected source ray can be calculated using the inverse square law. The image source method works by algorithmically finding all viable rays and their resultant image sources, then summing each of these signals to calculate the total signal at the receiver. For first order reflections, this is all that is required, however for higher order reflections, secondary image sources need to be calculated from the primary sources, then tertiary from secondary, etc [112, 113].

Figure 2.7b demonstrates a case where a resultant image source is not valid (S_B). As before, a source S reflected in surface B would form the image source S_B. However, as the path between the receiver and S_B will never intersect with surface B, it is an invalid source. An image source that is not visible to the receiver is also invalid. By this, it is meant that a path cannot be drawn uninterrupted between the two due to the geometry of the space [111]. For simple rectangular spaces, the image source method is at its most efficient. Since all of the image sources are calculated algorithmically, additional computations are required to check ray paths of collisions to guarantee that an image source is actually valid.

This method is simple and incredibly accurate, providing exact solutions to the wave equation for favourable geometries. But as the number of orders of reflection increase, and we account for the computational overhead regarding ascertaining visibility and validity of an image source, this method becomes less viable [114].

2.7.2 Geometrical acoustics: ray tracing

The technique of ray tracing is used predominately in graphical rendering, modelling light, shadow, and their behaviours in a scene. In geometrical acoustics, the source of a spherical wave can be represented by a set of lines originating at one point [115]. For an omnidirectional source, the direction of these lines is uniformly distributed about the point, however, we can also model directional sources too. These discrete rays radiate outwards, propagating at the speed of sound for the environment, obeying the laws of geometric acoustics, which in turn, rely on geometrical optics and Snell's law - the reflection edge is equal to the edge of an incident wave [116]. Each ray is traced until it drops below a certain energy threshold, which can be set pre run, and will be affected by distance to sound source, reflections off boundaries, and the propagation medium. The resultant sound can then be calculated by placing a finite-sized observer within the space and measuring the properties of all incoming rays to that observer. Figure 2.8 demonstrates a simple version of this.

In the past, this technique would have been best suited for applications where a longer

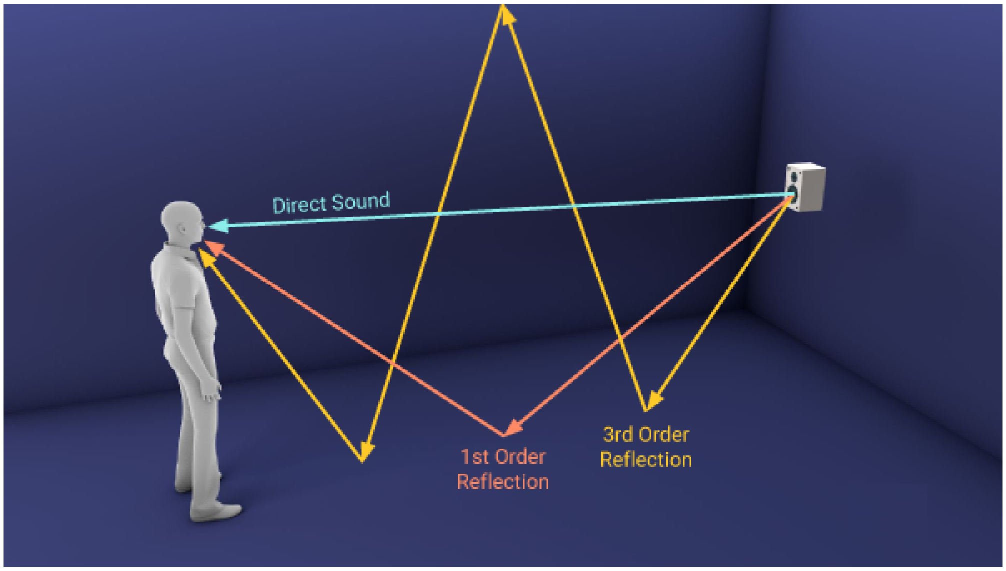


Figure 2.8: A simplified diagram of early reflections in geometric acoustics [3].

render time could be tolerated, as a more accurate render takes a great computational cost. However, thanks to advancements made in commercial game engines in recent years, real-time ray tracing has become a standard on newer graphics cards, allowing real-time rendered media to be more feasible.

Unlike the image source method, ray tracing is not limited to specular reflections, and is able to model diffuse reflections by emitting “bursts” of new rays at other surfaces when appropriate [105]. However, modelling sound as a finite number of rays described only by their power can cause a few problems. If not enough rays are emitted, some important paths may be missed when calculating the solution. Secondly, frequency dependent effects such as diffraction and interference are not accounted for when only power is considered [117]. Despite this, most virtual reality applications do implement a ray-based method as there are publicly available plug-ins created by Google and Steam (Resonance Audio and Steam Audio).

2.7.3 Resonance Audio

Google's Resonance Audio plugin is a popular geometric-based simulation method that is implemented frequently in the 3D game engines Unity and Unreal Engine [118]. Unlike other similar plugins such as Steam Audio, it also supports iOS and Web development, with native support for other game environments such as Android Studio. With Resonance Audio, it is possible to simulate key effects that help improve immersivity for a listener and immersive systems, for example, source directivity, occlusion, absorption, and reflection. This can be achieved in real-time, allowing for dynamic objects to also be included in occlusion calculations.

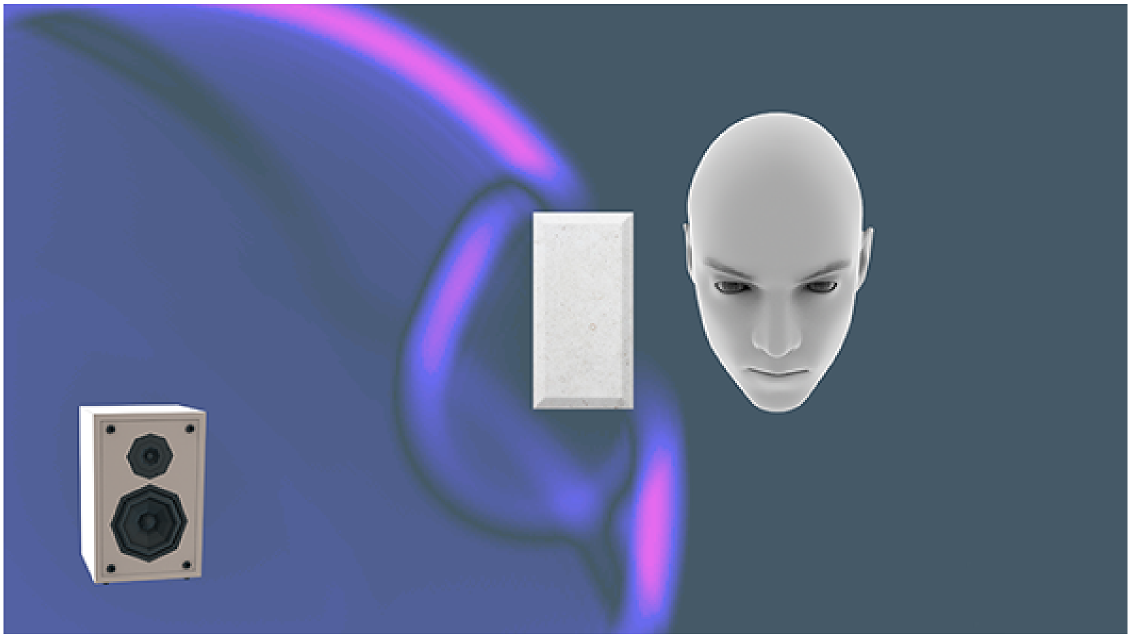


Figure 2.9: A simulation of how real sound waves travel for use in Resonance Audio, highlighting occlusion effects [4].

Before running the software, geometries in a virtual environment can be assigned acoustic properties based on the material allocated to them. Other factors such as reverb gain and brightness can be controlled through values assigned to the component. Then, through the use of probes that can be placed throughout the virtual environment, Resonance Audio pre-calculates the acoustic properties of the space by firing rays from these probes. These geometric rays calculate beam paths and the properties of the room, which are then implemented in run-time to modify any sound sources in the space to sound as though they are in the environment described by the material properties of the surrounding world.

These acoustic environment properties can be baked into the program in advance or can be calculated on the fly. In addition to the above properties, Resonance Audio also features sound source and listener directivity through simulation with HRTFs. This directivity is adjustable, with the ability for a user to change width, and patterns including circular, figure-eight, and cardioid [4].

2.7.4 Geometrical acoustics: beam tracing

Beam tracing is very similar to ray tracing, but instead of propagating rays throughout the scene geometry, a set of pyramidal beams are created at the sound source emanating in all directions [105]. These beams are extended outwards and tested against obstructing surfaces starting closest to the source. When a beam and object intersect, that beam is cut to remove any volume occluded by that object. Reflection and transmission from these objects can then be modelled [111]. Each intersection between a beam and surface can be thought of as an infinite number of rays interacting with a surface, which prevents the sampling effects that can occur with ray tracing. However, as each beam could be reflected and obstructed by multiple 3D surfaces, beam tracing requires relatively complex geometric operations when compared to ray casting or the image source method [111].

2.8 Recreating physical environments aurally

Through a combination of the aforementioned methods to create both natural visual and acoustic virtual spaces, there have been multiple studies exploring where these techniques overlap.

2.8.1 Virtual immersive audio experiences

In a 2017 paper titled *Virtual Reality in Church Acoustics: Visual and Acoustic Experience in the Cathedral of Seville, Spain*, a 3D model of a cathedral was created and integrated with realistic pieces of sound to make a virtual reality environment to explore quality of soundfield perception [5]. This study did not utilise an HMD and instead opted to project the visuals onto a screen. The aim of this study was to ascertain whether visuals, sounds, or what kind of sounds influenced immersion the most in a church virtual environment. It has been argued by Algargoosh that the physical measurements of architectural acoustics fail to precisely reflect the actual experience as it is hard to recreate majestic spaces authentically [119]. Even so, virtual reality experiences tend to focus on the visual side of immersion, but in a more acoustically reverberant environment such as a church, the audio is a very important factor.

To realistically recreate the acoustics of a worship building, there are four main tasks to complete:

- Measuring the acoustics of the space using a comprehensible set of RIRs
- Acoustic modelling and simulation to generate reliable RIRs
- Generating auralisations from simulated and real responses

- Modelling the interior of the space to round off the immersive experience, supporting the auralisation.

After registering the room impulse responses, simplified models of both the Cathedral of Seville and the Royal Chapel of the temple were created by using SketchUp 3D modelling software. Figure 2.10 shows these virtual models. The acoustic characterisation of the models was based mainly on absorption and scattering properties of surfaces. This was considered finished when the simulated reverberation time in frequency band, spatially averaged, differed by less than 5% from the true measured values.

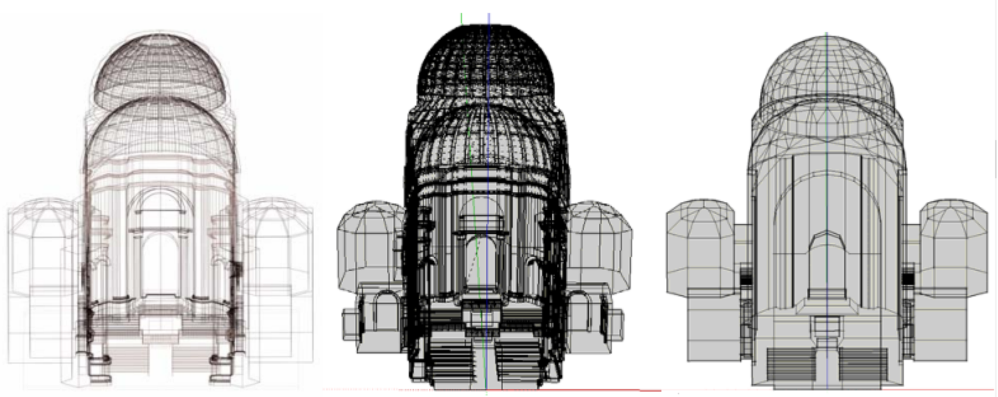


Figure 2.10: Simplified virtual models of the Cathedral of Seville and the Royal Chapel. Geometrical model before simplifications (left); simplified visual model (centre); and simplified acoustic model (right) [5].

The auralisations were tackled next. Álvarez-Morales et al. describe how essential Binaural Room Impulse Responses (BRIR) are for generating auralisations as they can be depicted as filters due to their incorporation of source directivity, HRTFs, and equalisation for the sound-reproduction systems used. Simulated BRIRs were convolved for three anechoic signals: a bible speech; a cello piece; and choral singing. To assess the accuracy of the generated auralisations, listening tests were created for 27 subjects. The participants were played 3 stimuli (modelled and real recordings) and had to discern which 2 were identical. Only

55.6% of participants could recognise which were the same, indicating that the simulated RIRs reproduced the real acoustics well.

Figure 2.11 shows the 3D visual model created for VR. This was achieved through using the detailed geometric models from Figure 2.10 and real photos.

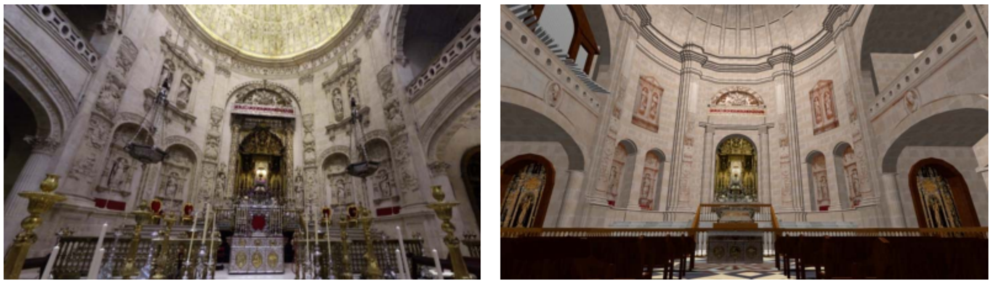


Figure 2.11: Photo of the Royal Chapel of the Cathedral of Seville (left) and the corresponding virtual VR model (right) [5].

The study was then split into three parts: where listeners were given only the auralisations and no visuals; where the participant received only visual stimulus and was asked to comment on things like lighting; and a complete virtualisation, including both visual and acoustic aspects, which was just the first procedure with visuals. The visuals were presented through a projector and sounds were emitted through headphones.

The perception questionnaire showed that the perceived reverberation of the space significantly changed with the stimulus, whereby it was acceptable for music and choral singing but very reverberant for word transmission. In comparison, when asked to rank the acoustics of the space after each sound piece, 91% of subjects classified the acoustics of the chapel as *good/very good* for choral and instrumental music, whereas 75% of subjects ranked it as *poor/moderate* for speech. The difference in responses for the combined part of the testing compared to the 'blind' aural testing were minimal, suggesting just how important audio is for virtual spaces as the addition of visuals did not unnoticeably affect the responses from subjects.

Other research groups have also explored immersive audio experiences in historical spaces. Selfridge et al. used off-the-shelf audio plugins utilised by standard game engines to recreate an 18th-century music room in VR [61] for example. The authors conducted listening tests in which participants compared music played in a real music room to the same music played in the virtual recreation of that room. Results showed that the use of off-the-shelf audio plugins such as Google Resonance and Steam Audio could recreate the acoustics of an environment "authentically", but that these recreations were less accurate than commercial room acoustic simulation software simulations.

Both these projects revolve around the user listening to music or speech that has been spatialised for an immersive virtual environment. The user is not an active participant in these experiences; they can only interact with pre-spatialised audio sources. However, research has also been undertaken to create interactive virtual environments, where the user's voice is convolved with an impulse response in real time.

2.8.2 Virtual interactive immersive performances

The creation of the Virtual Singing Studio (VSS) at the AudioLab, University of York, has been used to explore VR for real-time musical performance in multiple ways. A paper led by Gavin Kearney in 2016 outlines a prototype for an immersive and interactive virtual reality system for ensemble singing [6]. The aim of the project was to create a VR program through which the user would replace a singer within a vocal quartet performance at a specific location and be able to hear themselves singing with the group as though they were in that space. The quartet performed four times at St Olaves church in York. In each recording, one singer was replaced by a 6-piece GoPro camera rig and Eigenmike shown in Figure 2.12.

Sinusoidal sweeps were captured for each singing position in order to produce spatial impulse responses allowing for the virtual environment to have the correct acoustic. The

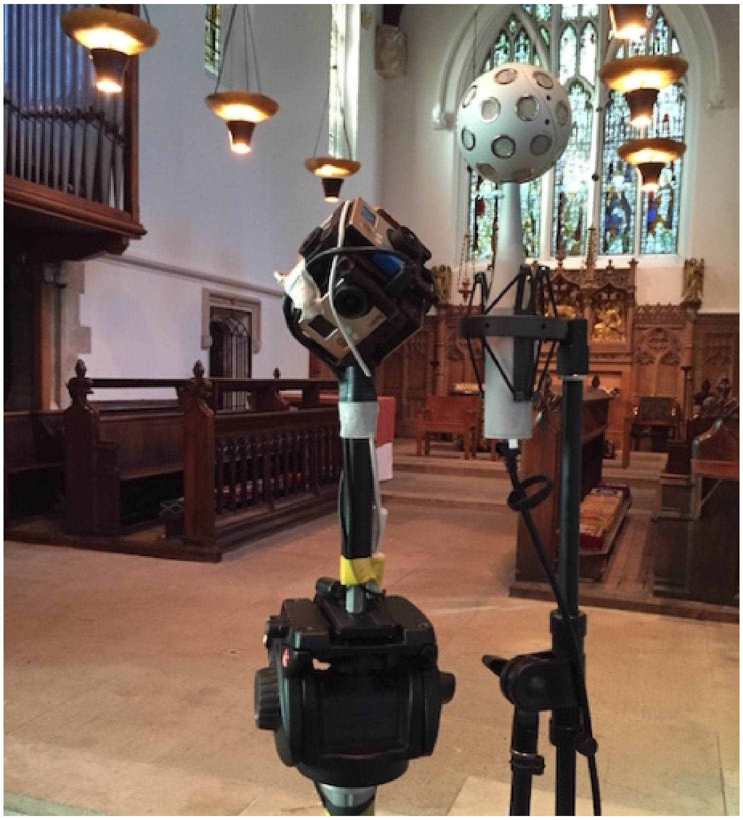


Figure 2.12: GoPro rig and Eigenmic set-up for singer replacement [6].

audio recordings are then put through a post-production procedure involving encoding into 3rd order Ambisonic files using the MH Acoustic EigenUnits plug-in for REAPER. Unity was used to create a skysphere with the footage from the camera, and REAPER was used to line up the new audio and visuals.

A tool called Max/MSP was used to play back the 2D 3rd order Ambisonic audio files [78], as it can be used to process real-time audio inputs and playback files whilst communicating with Unity. The full set-up for the experiment can be seen in Figure 2.13.

While in the centre of the speaker array, each singer was equipped with a DPA microphone. This allowed the singer's voice to be captured so they could hear themselves with the other three pre-recorded singers as though they were part of the original quartet recording.

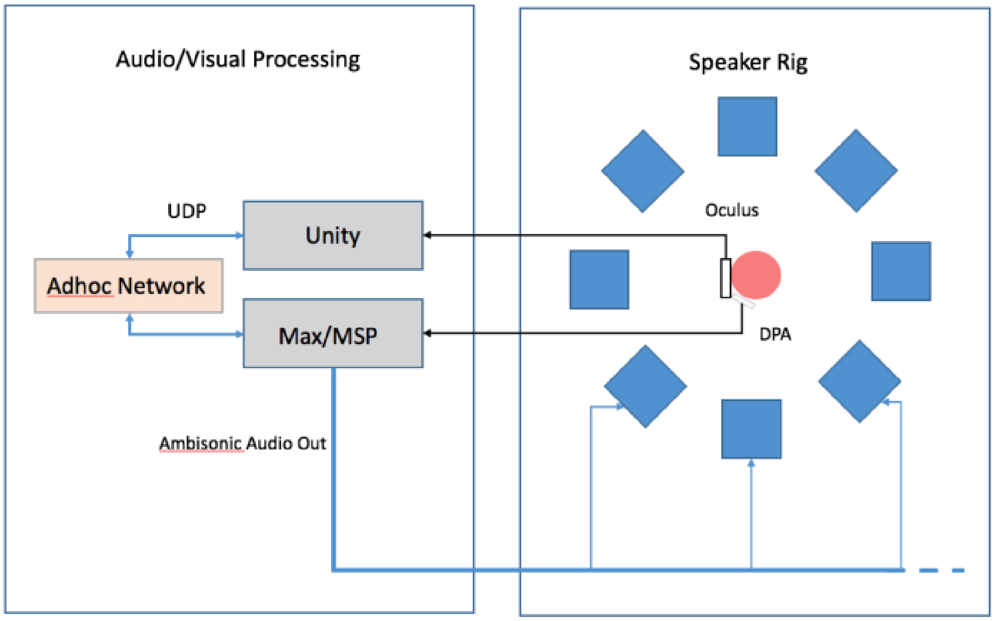


Figure 2.13: The set-up for the virtual reality experience [6].

A unanimously positive response was received by the team from the singers. All found the experience to be immersive despite the Oculus headset only providing a 100° field of view, and felt there was a lot of potential for the system as a tool for rehearsing and practice.

A second study titled *The Hills are Alive: Capturing and Presenting an Outdoor Choral Performance for Virtual Reality* from the 2019 Audio Engineering Society (AES) 2019 Conference on Immersive and Interactive Audio also explored using VR for performance, but from a different perspective [120]. In this project, a choir performed outside in the Lake District to create a VR experience which allowed the general public to become part of the performance in a virtual choir.

The mic and camera setup were the same for this project, but instead of replacing one person in a quartet, the camera was added to the centre of each voice type one at a time, for example, the sopranos, then altos etc. Rather than moving the setup into each group, the groups rotated around, each taking a turn surrounding the mic. An audience position was also added, as well as a choral improvisation camera for a different song.

The decision was made to put the camera at eye height, causing the microphone to sit at a lower point than the average height for a pair of human ears as it was decided the camera being at eye level was far more important for immersion. As this experience was to be presented to the general public in a museum, the surround sound eight speaker setup could not be used. Therefore, for this setup, the audio had to be converted to a 2-channel (stereo) version for headphones.

The 32 raw audio channels were encoded to 3rd order Ambisonics with `EigenUnit-em32-Encoder` VST from *mhacoustics* [121]. 3rd order ambisonics were implemented as it has been shown in multiple studies that as the Ambisonics order increases, as does the perceived spatial resolution. This means that the accuracy between different sound sources, or choir members is improved [122–124].

The 16 encoded ambisonic channels were then decoded to create a 26-point Lebdev grid, which creates a virtual speaker array. This is then binauralised for headphone listening. In theory, this allows the user to experience all of the individual multichannel audio channels, making the soundscape more representative.

Once again, a DPA microphone was used to allow the user to hear themselves back through the system and integrated with the pre-recorded choir. Unlike the previous experiment, this project did not use real-time convolution to simulate the acoustics of the real space. This is because Great Gable has a lack of reflective surfaces, meaning that there is only a negligibly noticeable acoustic, and various extra noises from wind and tourist traffic make the capture of a useful impulse response unlikely.

This set-up was brought to Kenswick Museum, where members of the public could sign up to take part in the project. The participants were asked a set of 5 questions asking them about their VR experience, whether they felt immersed, and if they would use virtual reality again. The response from the users was mostly positive, with most participants feeling as though they were part of the choir at the summit of Great Gable. The lack of tactile feedback does play a role in limiting the immersion of the experience as users can hear wind but not feel it for example. The team commented upon these limitations but summarise that the experience as a whole was positive.

The two York projects revolved around placing a person in the shoes of either themselves or someone else during a performance as one of the performers. In comparison, a study led by Luke Reed focused on a listener exploring height perception while listening to an orchestra [125]. To do this, a 51-piece orchestra was split over 5 levels of height and recorded with multichannel spot recordings. To yield a greater sound quality, the performers close to the camera were given spot mics in addition to the use of an Ambisonic microphone. All 32 channels of audio from the ambisonic mic could be tracked simultaneously to REAPER.

The team only had access to one 360° camera: a Kandao Obsidian S, which is a 6K stereoscopic 360° camera, therefore the performance was recorded as multiple takes [126]. This was to make the final artefact more interesting as the participant could move to one of 3 positions around the hall. Using Adobe Premiere Pro, REAPER, and the Facebook 360 Spatial Workstation (FB360) [127], the audio mixing was performed. An individual ambisonic mix was created for each camera position before the full editing was finalised. The video will automatically cut between shots, so it is not down to the user to navigate the scene.

Listening tests were performed to find the right balance between the Sennheiser AMBEO microphone and the spatialised spot microphones to make sure the acoustics were not too jarring. During editing, jump cuts in comparison to fades and dissolves were found to be preferable for the user. However, jumps could cause unpleasant disorientation when occurring outside of rhythmic phases of the music. To combat this, not only were cuts made at the ends of phrases, but instruments playing identifiable, clear motifs were positioned close to the cameras to really help orient the user, as it gave them a focus point after the cut. Overall, this project demonstrated a large-scale orchestral performance that made use of height for a VR Orchestral performance in a novel way.

2.8.3 Reverberation and perception

Much of the literature explored in this section so far has revolved around the recreation, or recording, of reverberant spaces. Touched upon briefly already in the work by Álvarez-Morales et al., the effect of perception of reverberation shapes our auditory experiences in various environments [5]. Understanding the perception of reverberation has been a subject of extensive research, incorporating both psychoacoustic experiments and computational modelling. The perception of reverberation relates the actual intensity of a sound source in a reverberant room to the perceived intensity [93].

A notable study by Girón et al. aimed to investigate the unique acoustic characteristics and the influence of architectural design on the subjective experience of reverberation in seven Spanish cathedrals [128]. Using sine-swept signals, RIRs were obtained for each space. Recordings of two musical pieces, one female voice, one male voice, and one choral piece were recorded in an anechoic space and through REAPER and Matlab were convolved with the RIRs and virtually placed a set distance away from the listener's position. During the study, participants were first presented with 15 pairs of stimuli and asked to indicate which stimulus was more reverberant in each case. After a short break, the participant is presented with auralisations of all five audio sources are presented to the listener corresponding to two of the seven cathedrals. Questions surrounding reverberation, intelligibility, and overall acoustic quality of the sound field of the cathedral are presented to be assessed, with participants rating their answers on a scale from 1 to 10, where 1 corresponds to the lowest value.

Statistical analysis of the survey showed that all listeners were able to identify the most reverberant cathedral for all audio stimuli, however in other cathedrals, there was no equality between objective and subjective reverberation. This led the authors to conclude that for their study, reverberation times between 4 and 7s render the sensation of reverberation indistinguishable for listeners. Furthermore, type of audio stimuli affected how participants perceived the reverberation of a cathedral. For instrumental and choral pieces, perception of reverberation time was lower than for spoken pieces, with the reverberation of the female voice being perceived as greatest. Only when considering the spoken audio samples is there any consensus across participants in ordering cathedrals from most to least reverberant.

Research into the perception of reverberation in small spaces has also been conducted. An experiment by Kaplanis et al. was undertaken to identify the perceptual effects of acoustical properties of domestic living environments, in a stereophonic reproduction scenario [129]. Spatial audio rendering was used to capture and reproduce nine sound fields which were then played from a 3D loudspeaker array. These sound fields represent a range of possible acoustic scenarios in a domestic environment, with varying room sizes and materials present. Three audio excerpts were recorded and convolved for comparing the various acoustical conditions; a piece of music, and percussion piece, and a speech.

Ten expert assessors evaluated the nine acoustical conditions by identifying their own perceptual attributes and quantifying their perceived sensations for the presented sound fields. They were tasked with discussing the perceived differences between the acoustical conditions, and the common perceptual constructs were identified and validated against the phsyical properties of the sound fields. The analysis of the results indicated that domestic rooms with a lower reverberation time are preferred, with a critical value close to the recommended mean value for reverberation time in a domestic environment existing. Above this value of 0 4s, preference for a space degrades. A commonality of these studies is that they aim to combine and compare physical properties of a space to hedonic responses provided by people, linking psychoacoustics with mathematical properties. This helps to quantify opinion with numerical data for future studies.

More specific in-depth research has been conducted on the perception of reverberation and its relation to binaural parameters. Specifically, the significance of early and late reflections for speech perception within reverberant environments [130]. The spatial characteristics, timbre, volume, and perception of the sound is affected by changes to the above reflections [131]. Early reflections play a crucial role in localising the source of a sound, providing listeners with cues about the room's size, shape, and their own position and orientation within it [132]. The early reflections that arrive back to the listener in the first 50ms after the direct sound are not perceived separately. Instead, they are integrated for directional cues.

Weak late reflections can result in a dry sound, while excessive late reflections lead to confusion and reduced intelligibility. However, when at an appropriate level, these late reflections can enhance recognition accuracy and sense of space [130]. Research on the importance of specific portions of an impulse response for different contexts has been investigated by Gölzer and Kleinschmidt [130]. They assessed the impact of early and late reflections on automatic speech recognition accuracy, finding that through the elimination of harmful late reflections, early reflections within a certain critical delay time can carry useful information, thus contributing to automatic speech recognition accuracy.

From this research, Mi et al. presented a study on the perceived importance of multiple acoustic factors of BRIR on perceptual reverberation [93]. BRIRs were generated for three different rooms and were convolved with a dry speech signal. Four reverberation parameters were altered for each room; Initial Time Delay Gap (ITDG), Forward Early Reflections (FER), Reverse Early Reflections (RER) and Late Reverberation (LR). These were convolved with the speech signal to create multiple audio conditions for a listener. Participants were tasked with finding the threshold at which each reverberation parameter was perceived as different from the original convolved speech signal. The average perceived impact threshold of each parameter for all twenty participants was then calculated [93].

The results of this article show that the removal of FERs did not have a significant impact on the perceived reverberation of the speech signals. The effect of Late Reverberations was hard to distinguish and was determined to not be serious as only when significantly large amounts of these Late Reverberations were removed was the perceptual reverberation affected clearly. However, it was found that the removal of RERs and extension of ITDG had a clear impact on the perceptual reverberation of the speech signals [93].

The experiments discussed in this subsection all have one thing in common; they all involved listening or interacting personally with performance in some variety. A mix of speeches, musical instruments, and singers have been utilised in both real-world recordings, and in virtual recreations of these real-world spaces through various audio spatialisation techniques. The explored literature has gone on to influence future chapters, sparking the initial ideas surrounding where novelty in this field lies.

2.9 Summary

This chapter opened with a brief introduction to virtual reality, and how one may recreate physical environments virtually. This was followed up by sections about immersive audio, as well as developments in spatial audio systems and auralisation of real environments. Sections 2.5 and 2.6 explored impulse responses; how they are recorded, measured, and convolved. Computational acoustic modelling was investigated next, detailing prevalent methods and techniques. Finally, previous research around virtual immersive audio experiences were described, exploring listening, performing, and speaking in virtual acoustic environments. Topics which these experiments did not investigate were highlighted to help indicate potential new areas of research for novelty within this thesis.

Accessible real-time room acoustic simulation is an area of interest that continues to grow, with multiple research teams investigating how one might integrate the acoustics of real-world environments in virtual experiences. Whilst there have been prior studies researching the thresholds for which acoustic parameters of a convolved impulse response are still perceived as correct when compared to a reference audio sample, for example, Mi et al., there have not been any studies combining this with real-time simulated performances [93].

The implementation of the studies conducted below differ in a number of ways from those outlined in Section 2.8. They include further investigation into the use of tools such as Resonance audio for the spatialisation of audio in immersive virtual reality environments, expanding upon the work of Selfridge et al. [61]. The work also expands upon the groundwork for the creation of immersive singing experiences and immersive singing technology as explored by researchers at the University of York. These studies also benefit from experimental results collected from frequent VR users, which is rare in studies of this nature.

The following chapter describes the first of three experiments investigating what is required to immerse users fully in a virtual environment. These chapters explore how the grey area between realism and authenticity can be exploited to better represent real-world aural environments virtually, implementing acoustic models that are directly based on a real-world environment.

Chapter 3

Effectiveness of Simple Audio Recreation

> *"There are many, many nouns for the act of looking – a glance, a glimpse, a peep – but there's no noun for the act of listening. In general, we don't think primarily about sound. So I have a different perspective on the world; I can construct soundscapes that have an effect on people, but they don't know why."*
>
> - Walter Murch, *Walter Murch: Searching for the Sound of the God Particle [133].*

As virtual and augmented reality systems develop and improve, the interest and demand for further applications grows. Although videogames could be considered the most developed area in general, exploration into use in research and entertainment over the last few years has caused an explosion of new experiences, for example from immersive 3D audio and video broadcast streams of live performances [134] [135] to VR in therapeutics [136] [137], and even virtual fitness classes [138] [139].

With all this growth, visual virtual interactions have improved rapidly, with real-world physics interactions, like the illusion of weight for a virtual object, even appearing in the latest VR games [140] [141]. However, audio plays an important part in convincing the user that a virtual space could be real and, although there has been exploration into improving immersion through advancing audio techniques [55,142–144], more needs to be done to bring audio capabilities to the same convincing level. Techniques such as individualised HRTFs for spatial audio in VR allows experiences to be personally tailored for a user. Using different speaker layouts for video games which more accurately convey spatial information can improve how engaged the player is. Real-time acoustics simulation through the convolution of audio inputs allows for aural interaction with a virtual environment in real time. These evolving digital technologies give the potential for virtual spaces to imitate real ones with increasing authenticity.

We experience spaces through not only seeing, but also by listening. Two aural components contribute to a listener's associations with a particular space: its unique sounds and its characteristic acoustics. For example, a rural outdoor space may have the sound of birds, the rustling of leaves and grass and wind. And outdoor spaces also have the distinct acoustics resulting from movement of reflective surfaces, thermal refraction and air turbulence. Every space, every environment, has an *'aural architecture'* and these aural attributes of physical spaces have contributed to human life throughout history [145]. The sounds of the world around us, whether noticed or not, create this symphony of life that we associate with specific spaces.

Achieving a synthetic aural environment that is an accurate representation of the natural world for which our senses evolved is not a simple task. There are many subtle audio cues in the world around us that help to dictate our opinion of a space regardless of what is being visually presented. The reverberant spaces of chapels and cathedrals for example exude reverence and awe, Greek amphitheatres enable sounds to be heard by as many people as possible, while lavish furnishings of expensive restaurants help to dampen sound, which can establish a feel of privacy and intimacy [146].

This chapter describes a project designed to explore whether regular VR users correctly identify audio recordings that were recorded live in a real, physical space over a virtually rendered one when the virtual audio environment has been created utilising "off the shelf" open source tools. Also discussed is whether or not these virtually rendered spaces are deemed clearly fabricated, natural, or too good to be true. The chapter begins with a brief discussion about the motive and rationale behind the study and what was hoped to be gained from it. Section 3.2 discusses the methodology of the experiment, with a description of how the visual and audio recordings were recorded, and what techniques were utilised to create the virtually rendered audio sources. The results of a questionnaire of both participant responses and experience are recorded and discussed. The chapter concludes with an analysis of the results, ideas for improvements and similar study ideas, and a discussion of future research using the techniques discussed.

3.1 Real-world audio experiences can be transparently recreated in Virtual Reality

The goal of this study was to explore whether the acoustics of real-world spaces could be recreated authentically and in such a way that, when compared to live recordings, were either indistinguishable from them, or natural feeling if different. For this project, the viability of a simple and cheap impulse response creation technique was explored, focusing specifically on utilising a balloon pop. This was partially due to available resources at the time, but also as it was deemed interesting to investigate this method to ascertain if this simple process for calculating the impulse response of a space would be suitable for future research in this report.

Within this study, it was also advantageous to explore how the acoustics of an environment would affect the imitation of a space, or whether it would make it easier or harder to ascertain the artificial acoustics from the live recordings. Would reverberant spaces that are designed to be echoey be easier or harder to imitate authentically? Would open outdoor spaces with minimal reflections require an IR measurement to start with? An additional facet to explore would be whether different audio sources would elicit different results when considering the above questions. For example, would there be a difference between distinguishing voices compared to musical instruments? Would the sounds of an orchestra in an open field be conflicting for people as the audio experience contrasts with what may be expected from the visual space?

Finally, as will be discussed in Section 3.5, all participants in the final study were regular virtual reality users, or generally familiar with VR. Research in audio VR rarely benefits from participants who are used to VR, limiting some of the questions a researcher is able to pose. In studies reviewed as part of this work, there were very few frequent VR users among participants, [120] [57] [147] to highlight a few. Here, this research is able to probe high-end expert listener bases who are comfortable using a VR headset.

To investigate these questions, two acoustically different locations on Royal Holloway's campus were chosen; a reverberant indoor environment (the Chapel) and an open outdoor space with few surfaces for audio to reflect off (the Meadow). These locations were chosen mostly for convenience, but also for the availability to record and rerecord at them as required. It was also decided that a female speaking voice, a male speaking voice, and a snippet of orchestral music would be used to compare differences in audio stimuli in these spaces.

3.1.1 Concept and motivation

The motivation behind this study is simple: it was anticipated that recordings that were produced to sound as though they were recorded in a different environment would be natural sounding enough to the regular VR user to make them think that the recording was live, and if not, that it would be a decent enough impression of the space for use in further studies.

To explore this, studio recorded audio performances of a female voice recording, male voice recording, and an orchestral recording would be played in each physical space from a commercial speaker and recorded with a 360° camera and microphone. This recording set-up replaces the audience member in this scenario, and when formatted becomes what a person would see when watching the performances in VR.

By measuring the acoustical properties of the original physical performance environments, these studio recordings could be convolved to make them sound as though they had been recorded in the live locations, and virtually placed in the same location with regards to the stationary listener.

Therefore, a participant in the experiment could be given two recordings of the same audio stimuli in the same visual and aural location (either the Chapel or Meadow) to experience in virtual reality, where one of the recordings was virtually rendered to sound as though it was recorded live. This could be repeated for multiple audio sources to explore the full range and extent of both the convolution software and the physical environment, with the participant acting as the audience.

3.2 Recording in real acoustic environments

The intention of the study is to compare and contrast pairs of recordings; one where both the audio and visuals were recorded at the physical location, and one where the video was filmed at the physical location, but the audio has been added in post to the scene after convolving it. As such, the decision was made to pre-record all audio samples being used. These recordings were made to maintain consistency throughout the study, preventing the speaker from subconsciously adjusting their speech when recording in more reverberant spaces compared to less reverberant ones, as people tend to, for example, slow their speaking down in more reverberant environments [148–150].

To produce this “live recording” version of the experience (as it will be referred to from now on in this chapter), these pre-recorded vocal and instrumental performances would be played from an Anker Soundcore 2 portable speaker in situ at the two physical locations. As such, these are recorded “live” alongside the visual recording.

The “falsified recording” version of the experience (as it will be referred to from now on in this chapter) was created by taking the pre-recorded audio samples, convolving them with IRs from the physical locations, and placing them in the correct place in an artificial audio environment through REAPER. These recordings are created without playing them live at the real location, hence their description as “falsified”. These processes will be explained in further detail in Section 3.3.

3.2.1 Audio sample recording

With a recording studio not available, a male and female voice were instead recorded in the middle of a large, open field at night to pre-record samples with minimal sound reflections and external noise [151] [112]. These recordings were made with a Zoom H3-VR microphone and converted to a mono recording with REAPER: a digital audio workstation [152]. Specifically, the inbuilt Mono (downmix) function was used. This was done as these recordings were to be utilised in two ways; played from a speaker live at the location and placed into a virtual sonic environment as a localised audio source. Therefore, a mono recording was required.

Both the male and female performers recorded the same short speech at each location, describing some features of each space. The script for the Chapel recording is as follows:

> *Welcome to the Chapel at Royal Holloway, University of London. Take a moment to look around and above you. The Chapel is full of beautiful gilded artwork. Many societies and faiths use this space throughout their time here. This includes a tradition of daily sung morning services.*

And for the Meadow recording:

> *Welcome to the Meadow at Royal Holloway, University of London. Take a moment to look around and above you. Ahead of you, the tops of the Founders Building can be seen; one of the two original buildings at this university. Built in 1879 as a university for women, it now holds around 10,000 students studying a variety of subjects.*

Mozart's Serenade #10 in B Flat was chosen as both the reference track to check the validity of using a large open space for recording the dry vocal performances, and as a third recording for participants to compare in the study. This piece was chosen as we hear a wide range of frequencies and timbres, as well as a wide dynamic range, within the first 30 seconds of the piece; allowing for a more dynamic exploration of the acoustic properties of both spaces. Classical music is often recorded with extra attention to recording quality, resulting in a reasonably dry recording when replayed from a speaker [153]. The specific recording used for this project is referenced here [154].

To ascertain whether the spoken recordings were dry enough to not be adding their own extra reverberation to the scenarios, this orchestral clip was played and recorded at the recording location of the pre-recorded voice samples. This recording was then compared to the original orchestral file. Figure 3.1 displays these comparison spectrograms. Due to their similarity, this demonstrated that the voice recordings from the open field could be assumed to have negligible extra reflections, making them suitable for further manipulation within the study.

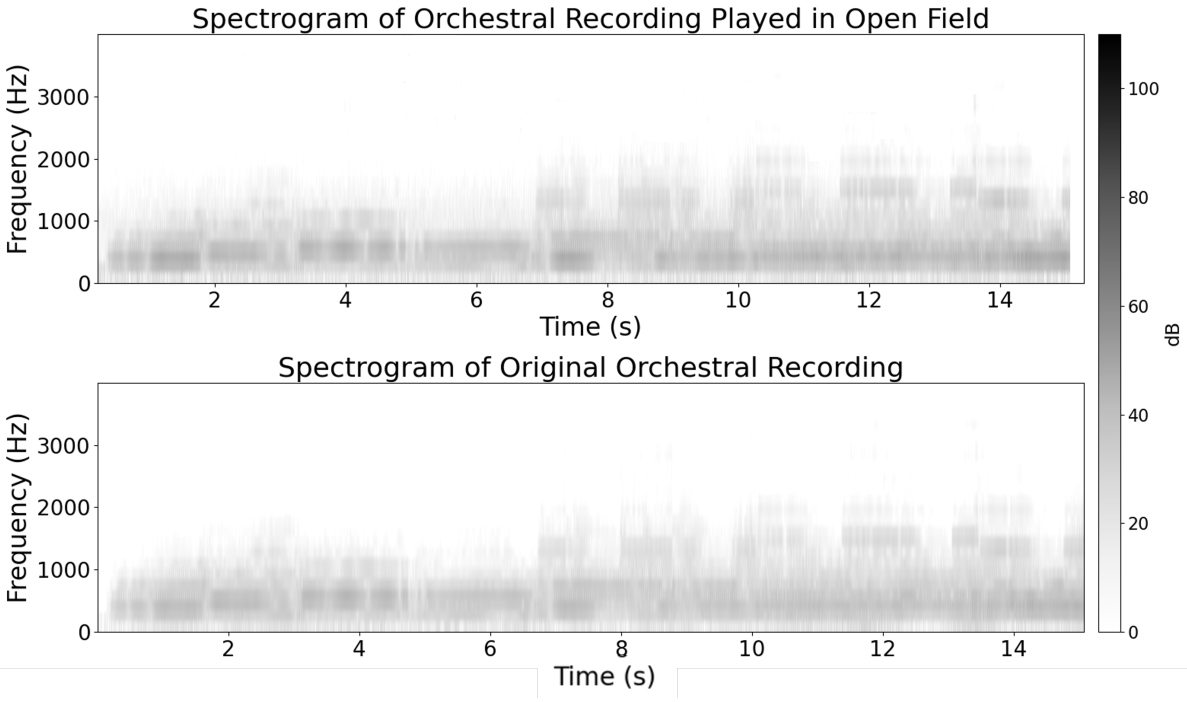


Figure 3.1: Two spectrograms comparing the sample recording location's acoustics with an original clean recording of an orchestra.

The frequency response curve of the portable speaker is given in Figure 3.2 [7]. Ideally, the speaker used would have a flat frequency response that does not amplify or attenuate any frequency ranges. As can be seen, this speaker does not have a flat response, with the mid frequency ranges having a higher amplitude that the low or high ones. The use of this speaker was required due to the circumstances at the time, and it is possible that the frequency response of the speaker used will have an effect on the perception of the output. Investigation of the effect of a speaker's frequency response on the production of sound via convolution with room impulse responses is beyond the scope of this work and is left as a future subject of research.

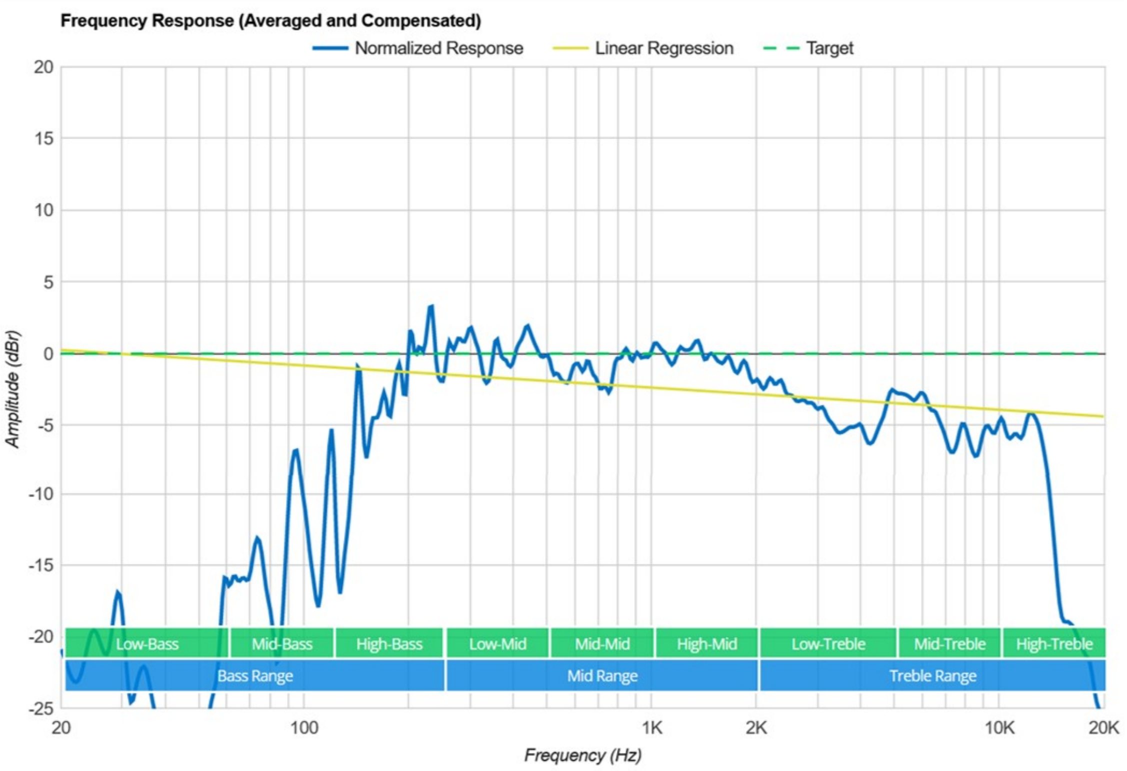


Figure 3.2: A graph to show the frequency response (averaged and compensated) for the Anker Soundcore 2 Speaker [7].

3.2.2 Impulse Response measurement

To create and measure the IRs of the two recording locations, a balloon was popped and recorded. An impulse response consists of a single, impulsive sound with a large amplitude. Balloon pops exhibit a short, Dirac-like impulse sound, therefore it is common practice that they are used in measuring impulse responses of spaces. A balloon burst is a cheap and easily available resource, which was why it was specifically chosen for this study, however as discussed in Chapter 2, there are other methods in which to measure an IR [155]. The exact locations of where the balloons were popped and recorded were noted, and these locations were then re-used when recording the "live recording" at each location.

3.2.3 Scenario recording set-up

The downmixed sample voice and orchestral recordings were downloaded onto a phone and played in the two physical locations through an Anker Soundcore portable speaker. These live performances were recorded with the H3-VR microphone and Go Pro Fusion 360 camera [80] [54]. The H3-VR is a first-order ambisonic microphone that can record in both Ambisonics A and B formats, with the encoding and decoding handled by the device if desired. As discussed in Chapter 2, A format equates to the raw signals outputted from the four capsules, as shown in Figure 2.3, which can be combined together to convert to B format.

The Go Pro Fusion contains four microphones, and with an on-board transformation algorithm, these raw audio recordings captured are converted into the ambisonic format. It also utilises 2 cameras with fish-eye lenses to create 5.2K spherical footage. These audio and visual recordings were combined to create the "live recording" half of the scenarios.

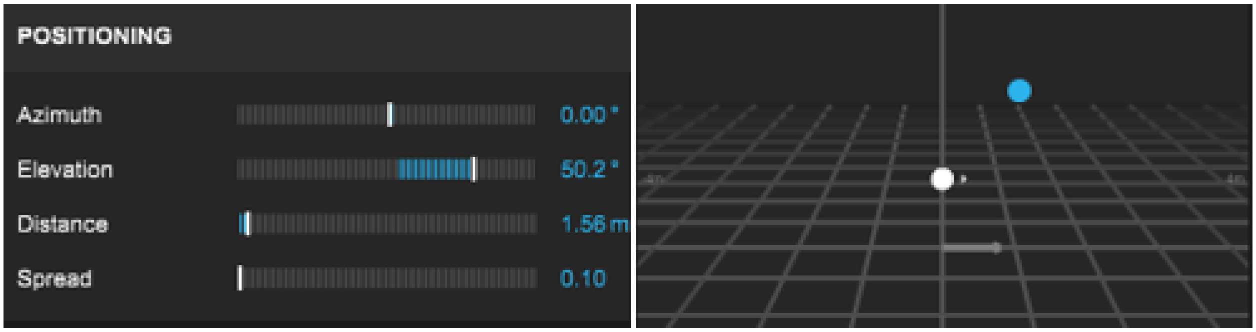
A further recording of just the background noises of the Meadow (birdsong, traffic etc.) was also made to be used in the "falsified recording" scenarios during rendering. As this study is looking at the viability of using the real IRs of a physical space and applying them to the spoken/played audio in a virtual space, this seemed like a fair way to ensure that outside factors, such as lack of bird song, would not affect participants' responses. No convolver can change the fact that there are no outdoor background sounds present in the "falsified recording", thus making the decision of which recording is live trivial without this. This recording was also made from the same physical place within the Meadow to align with the other visual and audio recordings.

3.3 Rendering

3.3.1 Audio rendering

Convology XT, a free convolution reverb plugin, was used to apply the impulse responses from the Meadow and Chapel to the sample vocal and orchestral recordings [156]. This plugin manipulates the recordings by convolving them with the impulse response to make them sound as though they had been recorded in a different location: where the IR was initially recorded. The user uploads an impulse response, feeds either pre-recorded or live audio through it, and it outputs the convolved recording. This was used to create the "falsified recordings" by playing the dry, mono pre-recorded voice and orchestral samples through the convolver.

Facebook 360 Spatial Workstation was then used to place these sound sources in the correct physical places from the perspective of the camera [157]. This is a software suite specifically designed for projects using spatial audio for 360° video and cinematic VR. By using the FB360 Spatialiser plugin, we can gain full 3D positional control. A down-mixed mono track of the convolved female voice, male voice, and orchestral recordings were placed at the correct distance and elevation away from the listener so that they lined up with the location of the speaker in the video recordings. This artificial audio scene can then be exported as an ambisonic file. Figure 3.3 shows the relevant part of the FB360 interface that was used to recreate the audio scenarios. The elevation was calculated with simple geometry calculations, as the speaker's height and distance from the camera was known.



POSITIONING

Azimuth 0.00 °

Elevation 50.2 °

Distance 1.56 m

Spread 0.10

Figure 3.3: Part of the interface of Facebook 360 Spatial Workstation. Shown is the placement of a mono track (blue sphere) at the position of the visible speaker system in the recorded visuals for the listener's (camera) position.

3.3.2 Video rendering

These live and falsified recordings were then combined with the 360° videos in Adobe Premiere Pro, a software which can handle and export ambisonic audio and 360° video [158]. To set up the work environment, a few settings need to be altered so that the ambisonic recordings are formatted properly. Firstly, when the audio is imported into Premiere Pro, its audio channels need to be modified. The Clip Channel Format needs to be set to Adaptive, with the active channels per clip set to the number of channels of the recording (e.g. 4 for first order ambisonics) and the number of audio clips set to 1. The Sequence Settings also need to be matched to the video and audio properties required. The width and height for the video settings can be found by looking at the properties of imported videos for the project.

Premiere Pro can also be utilised for adding spatial audio to YouTube videos, which can handle first-order ambisonic audio files.

Although the GoPro Fusion does record first-order ambisonic audio on board, the recordings are of a lower quality when compared to a dedicated ambisonic microphone, therefore the videos were exported from the device as stereo recordings to reduce the file size and export time. The audio was kept to assist with the alignment of the microphone's recording with the video, as at the start of each scenario, an impulse was produced to create a sharp spike in both of the audio recordings. This audio was then removed from the final mix.

Premiere Pro contains a feature that also allows you to check whether your audio and video recordings are aligned directionally correctly. Ambisonic audio recording effectively creates a static sphere of audio that is centred about the listener, or microphone, in the same way a 360° camera captures a sphere of video that is stitched together. This is easy to check in VR once exported as by turning your head to the right, you would expect sound sources that were playing in front of you to now sound as though they are closer to your left ear. However, exporting a video each time to check if the spheres are correctly aligned is very time consuming. Instead, this same effect can be achieved through applying the "Binauralizer - Ambisonics" effect in the Audio Track Mixer. This converts the 4-channel audio into binaural audio, and by virtually turning your head with the accompanying panner dial, it can be ascertained whether or not the visuals and audio are aligned correctly.

If they are not aligned, the "Panner - Ambisonics" effect can be applied to the audio files in the sequence, allowing you to pan, tilt, and roll the audio sphere to its desired orientation. Before exporting, the binauralizer effect must be removed to actually receive the desired ambisonic audio.

For the "falsified-recording" scenarios on the meadow, the bird song track recorded at the location was added to the mix to make it a more natural outdoor experience. These were recorded on the same day as the other recordings and were oriented correctly with the rest of the audio. Then, each pair of the live and falsified videos were randomly labelled 1 or 2 so as to not unconsciously bias the labelling process or form a pattern.

3.4 Method

3.4.1 Set-up of virtual environments

Vive Cinema was chosen as the software to manage the playing of the ambisonic videos on PC-based headsets as it is an open source project that is compatible with many head-mounted display (HMD) systems [55]. This is a lightweight, high-performance VR video player implemented by OpenVR and is fully open access. Initially released by HTC co-operation [159], researchers at the University of Parma adapted it for use with uploaded individualised HRTFs that can be switched between while playing files. A general HRTF was selected to use with this study as it was not possible to produce unique HRTFs for each participant under the circumstances.

Unity and Unreal Engine were also explored when creating this project but were ultimately not used in this study due to the functionality that Vive Cinema provided. Section 3.6.1 goes into more detail behind this decision and ways in which these programs could be utilised to improve the overall study experience.

3.4.2 Pilot study

A pilot study comprised of 10 participants was undertaken prior to the running of the final experiment. Each person took part in this pilot with the researcher present, and all used an Oculus Go headset with a pair of on-ear, Audio-Technica ATH-M50x headphones, as the

VR headset is fully portable. Participants were encouraged to stress test the experience and move their head in ways that they may not ordinarily. This did highlight some issues with the study. In one video, the audio clipped, then got louder when a participant tilted their head by a 90° angle with one ear pointing up to the ceiling. Participants also commented on how hard it was to mark their confidence out of 10, as that felt like too broad a spectrum. These responses were taken into account, and the study was altered. Participants were instead asked to rate their confidence in their responses from *Confident* to *Unsure*, with five available answers instead of 10. When it came to the unexpected audio effect on one video, the original Premiere Pro files were revisited and exported for a second time. This fixed the irregularity in the audio.

3.4.3 Further study adaptations

On advice from the World Health Organisation, the COVID-19 pandemic halted progress in this study, as it required a researcher to be present with the participant, which was illegal at the time when alterations to the study had been completed. The research project was revisited a few months later and adapted to be run remotely without a researcher present.

A Google Drive folder was created with all the components required for anybody with a VR headset to undertake the study. This included a document explaining all the steps required for setting up the experiment remotely. A folder containing all of the videos needed for the study was included for participants who owned a portable headset, as they would be able to download the videos straight onto the device. A pre-zipped copy of the Vive Cinema video player was included as well. Participants were informed that they could use whichever 360° video watching app they wanted, but that it had to support ambisonic audio. They were then instructed that if they did not know what this meant, or if they could not verify this, that they should use the attached video player. The videos for the study had already been saved to the correct location within the zipped file, so participants only needed to download and unzip the folder to begin. The videos were relabelled slightly, still maintaining their randomly assigned 1 or 2 for each pair, just to make the order in which to watch the videos unmistakable. It was then stressed how the study required the use of headphones or off-ear speakers to complete.

A link to the information sheet detailing the motivation behind the study was attached, along with a link to the consent form and question sheet. The question sheet itself provided instructions throughout the undertaking of the study. A question asking which VR headset a participant would use for the study was also added as a way to see if different headsets elicited different results. The questionnaire, Information Sheet and explanatory document can be seen in Appendix A.

3.4.4 Procedure

Participants were sent this Google Drive link that led to all the required materials and instructions to participate in the study. The root of this folder contained a document called "`READ ME FIRST PLEASE`", which provided all the instructions and technology requirements necessary to complete the study, for example, the necessity of headphones or off-ear speakers. For ease, relevant forms such as the information sheet and consent form were hyperlinked straight into this document.

The participants were lastly presented with a question sheet which provided a walk-through on how to take part in the study, instructing them to answer questions and when to put on and take off their headsets. Each participant was instructed to listen to and watch each pair of videos (the "live recording" and "falsified recording" for each location and audio source pairing), decide which one was the "live recording", and state how confident they were in their answer. These answers were submitted virtually through the sheet. At the end of the experience, participants were also questioned about whether the audio within the scenarios ever made them feel as though they were at the physical location, with space to leave written explanations of their thoughts.

3.5 Results

3.5.1 Participants

34 participants took part in the study, providing written informed consent, in accordance with the Royal Holloway Research Ethics Committee. Most of the participants were already familiar with VR headsets as they were invited to take part through four subreddits (forums) dedicated to virtual reality. A small number of participants took part in person as per government guidelines. Neither gender nor age were considered for this study as it was deemed inappropriate.

3.5.2 Questionnaire

Participants were asked a set of three unique questions per pair of sound sources (female voice, male voice, and an orchestral recording in the Chapel and the Meadow):

1. Of the two scenarios, which one did you think used the live recording of the location?
2. How confident are you in your answer?
3. Why did you pick these answers? (feel free to leave blank)

And then a set of questions inquiring about the experience as a whole:

1. During the Chapel scenarios, did the audio ever make you feel as though you were at the location?

2. During the Meadow scenarios, did the audio ever make you feel as though you were at the location?

They were also asked to state which VR headset they used to complete the study.

As discussed in Chapter 1 the definition of immersion is not consistent in the literature. Therefore, asking participants about whether or not they felt immersed in the space could have led to ambiguity in how participants interpreted the question, and also in their responses. This is why participants were asked to consider if the audio experiences made them "feel as though [they] were at the location", helping to establish whether the virtual experiences authentically recreated the equivalent physical space.

3.5.3 Questionnaire results

Which scenario used a live recording?

The results from Q1 can be seen in Figure 3.4. The results for all but one of the pairs deviate from a 50/50 result by no more than 9% (3 votes). The average of all the results shows a slight tendency for choosing the live recording (52.5% of total votes), but most of the individual comparisons were slightly in favour of the falsified recordings.

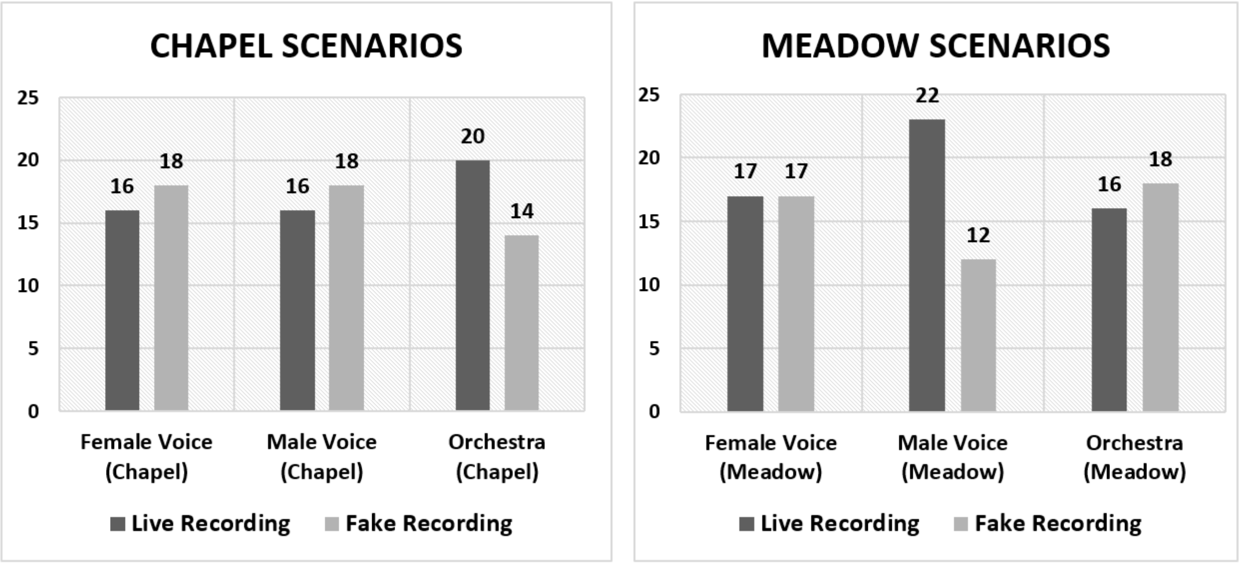
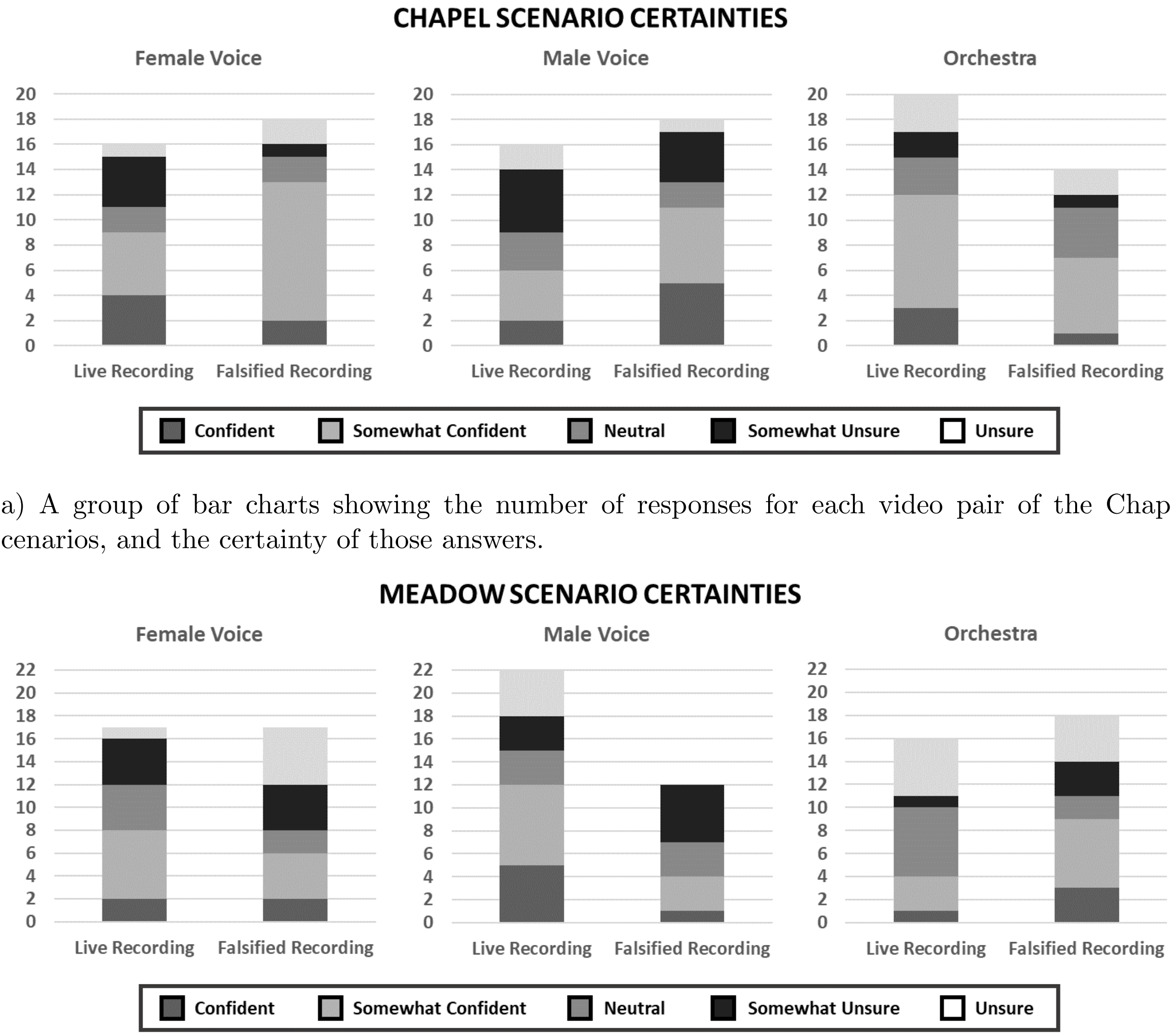


Figure 3.4: A group of bar charts which show the number of responses for the live and falsified scenarios for the question *Of the two scenarios, which one did you think used the live recording of the location?* Participants were choosing between videos randomly labelled 1 and 2 and did not know which audio was falsified.

How confident are you in your answer?

Participants were then asked how confident they were with their responses. For every pair of recordings, a mix of certainties were recorded. In general, most participants were more certain in their answers, responding with *Confident* or *Somewhat Confident* (57%) for the Chapel videos. However, for two of these three chapel scenarios, this majority confidence was placed in the falsified recording. Participants were less certain for the meadow videos (average of 42% *Confident* and *Somewhat Confident* responses), resulting in a more varied response across the 3 sound sources, with no real correlation between confidence and choosing the live or falsified audio.

There does not appear to be much correlation between respondents who were confident and respondents who were correct. For example, only 31% of those who responded with *Somewhat Confident* or *Confident* for the Orchestral Meadow videos were correct, compared



(a) A group of bar charts showing the number of responses for each video pair of the Chapel scenarios, and the certainty of those answers.

(b) A group of bar charts showing the number of responses for each video pair of the Meadow scenarios, and the certainty of those answers.

Figure 3.5: The results for the questions asking participants to choose which video they thought was the live recording and state their certainty in their answers for each pair of recordings.

to the other extreme of 75% correct for the Male Voice Meadow. However, even though more participants were certain in their answer for the Chapel on average, fewer of them chose the real recording. Whereas for the meadow scenarios, the lower average certainty resulted in a higher average percentage of participants choosing the real recording (54%). Figure 3.6 highlights some of the reasons why participants chose their answers, as well as the percentage of correct answers for both kinds of scenario.

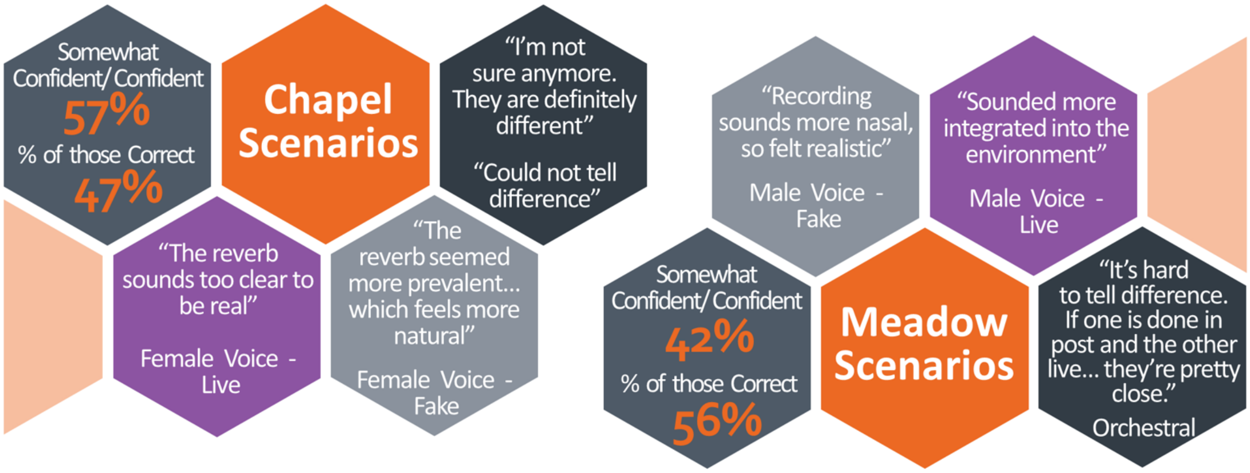

Somewhat Confident/ Confident
57%
% of those Correct
47%
Chapel Scenarios
"I'm not sure anymore. They are definitely different"
"Could not tell difference"
"The reverb sounds too clear to be real"
Female Voice - Live
"The reverb seemed more prevalent... which feels more natural"
Female Voice - Fake
"Recording sounds more nasal, so felt realistic"
Male Voice - Fake
"Sounded more integrated into the environment"
Male Voice - Live
Somewhat Confident/ Confident
42%
% of those Correct
56%
Meadow Scenarios
"It's hard to tell difference. If one is done in post and the other live... they're pretty close."
Orchestral

Figure 3.6: A showcase of a selection of the responses and reasoning for the choices made when listening to both scenarios.

Did the audio ever make you feel as though you were at the location?

Participants were asked to give their view on whether they felt immersed in both the Chapel and meadow scenarios, answering *yes*, *not sure*, or *no* for each location. Only four participants said *no* for the Chapel, with just 3 of those same 4 people saying *no* for the meadow. The full breakdown can be seen in Figure 3.7.

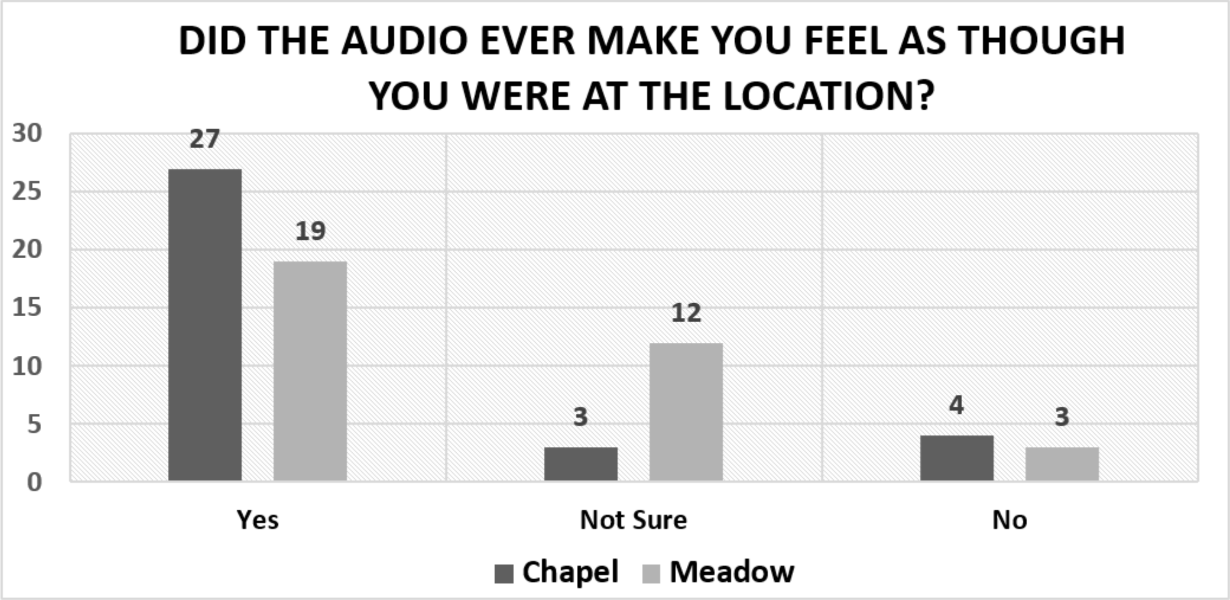


Figure 3.7: A group of bar charts to show the number of responses of *Yes*, *Not Sure*, and *No* for the question *Did the audio ever make you feel as though you were at the location?* for both the Chapel and meadow scenarios.

For the 7 people that replied *no* or *not sure* for the Chapel location, all but one person who did not feel like they could make a decision due to never being in a Chapel before, gave basically the same reasoning; regardless of the spatial nature of the sound, the visuals reminded them that they were just watching a 360° video. However, the majority of participants were incredibly positive about the experience, saying that the “richer echoes”, “directional aspect”, and “sense of space” helped to immerse them in the Chapel.

Participants were less certain when it came to the meadow, although people still found the outdoor experience immersive, mainly because of the bird song present. The meadow has a lot of wildlife, so there is loud bird song all year around. A few people mentioned that they thought “the less echoey space made it harder to feel real”, which is understandable. If you are even slightly familiar with reverberant spaces like a chapel or hall, you have an idea of what you expect to hear, whereas open spaces have less distinct acoustics due to, for example, fewer surfaces for sounds to reflect off of.

Bird song and wind help with outdoor locations, but as both the falsified and real recordings had these present, participants had fewer audio cues to differentiate the scenarios, as could be seen from the participant's responses. However, many of the people who replied with *yes* stated that the audio made it sound like they were outdoors, and that the "ambient noises in the background really helped".

3.6 Discussion

It can be reflected from the results of this study that real audio spaces have been successfully recreated virtually, therefore achieving the core aim of this project. On average, there was a slight tendency to choose the live recording (52.5% of all votes) over the falsified one, however, this almost even split suggests that the fabricated recordings held up when compared to the live ones. This can be seen in the results above, with the falsified recordings having the same or majority vote in 4 of the 6 scenarios (see Figure 3.7). Most of the participants who left reasoning for their choices noted the differences between the pairs of recordings when rationalising their answers. The scenarios that elicited the most responses stating that no differences could be determined was for the female voice recordings, with 16.7% of those who left a written response stating that they "both sound really similar", and more specifically for some that they "couldn't hear any difference between the two chapel recordings". None of these participants state that they can discern no difference in sound for the later audio sources, with one suggesting that this newfound ability to differentiate the recordings could be down to them "... paying a bit more attention... this time".

For the male voice recordings, only one participant noted that they could hear no difference, and this was just for the recording in the chapel, and 9.5% of participants who left a response for the orchestral recordings could discern no difference either. This demonstrated that the falsified and live recordings did not sound the same for most people. So, although real audio spaces were successfully recreated virtually to such a degree that participants believed the false ones to be the most natural 47.5% of the time, these spaces were not recreated perfectly.

In the most extreme result, the live Male Voice (Meadow) recording, potential reasons as to why there was a highly favoured answer have been explored. Using PRAAT, a phonetic analysis software, the average frequency ranges of the originally recorded female and male voices have been calculated [160, 161], as well as their higher harmonics. Gaussian white noise was played through the meadow balloon pop to determine whether there were specific frequencies or ranges of frequencies being amplified by the IR of the meadow. Figure 3.8 shows the signal created, as well as highlighted bands spanning 10% on either side of the average voice pitches.

In general, many of the signal's frequencies encompassed by the average male voice have a higher amplitude, which could affect the way the falsified recording would sound by artificially amplifying parts of the recording due to the Superposition Principle [162]. This could be why more participants than on any other comparison could determine the live recording over the falsified one.

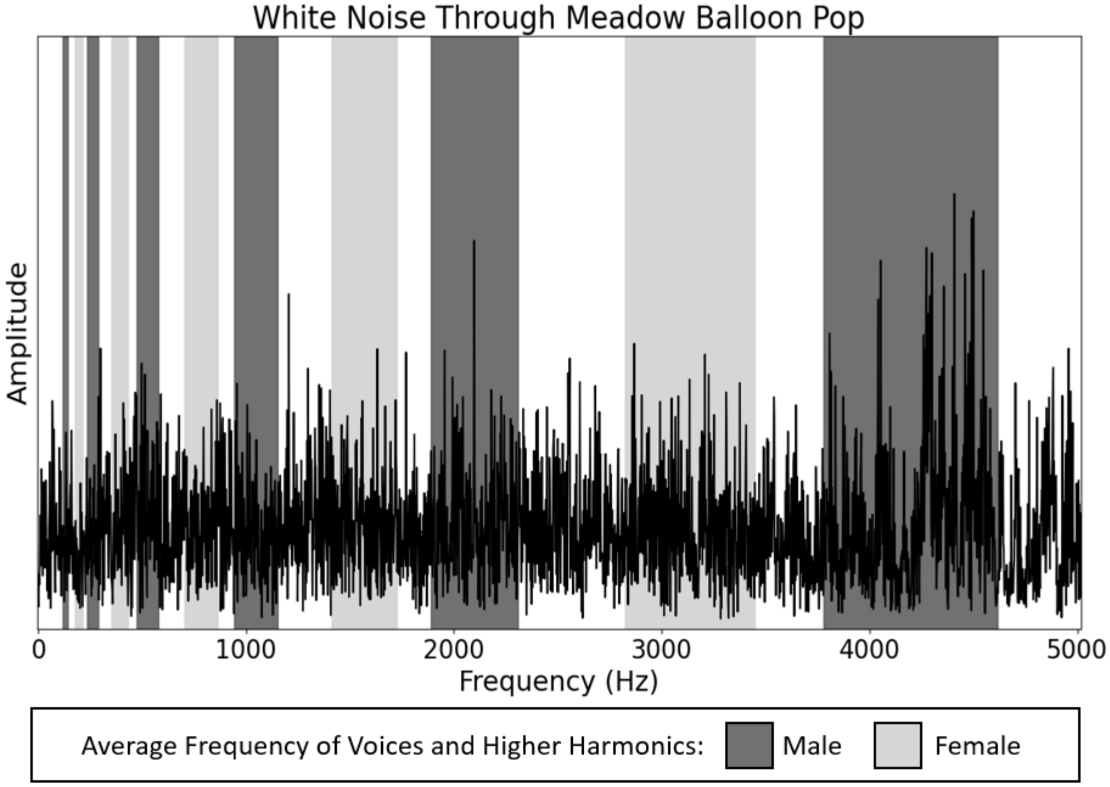


Figure 3.8: A graph to compare the amplitude of the frequencies created from passing white noise through the IR of the meadow with the average frequencies and higher harmonics of the female and male voices.

To determine whether the tendency towards selecting the live result for the Male Voice (Meadow) scenario was meaningful, a chi-square test was used. The chi-square compares frequencies obtained in the sample to those expected according to the null hypothesis (i.e., the inability to correctly identify the live recording over the falsified one, resulting in an even split of 17 votes to each) [163]. Using the chi-square formula,

$$\chi^2 = \sum \frac{(O-E)^2}{E}, \tag{3.1}$$

where χ^2 is chi-square, O is the observed frequency of a result, and E is the expected frequency of the result if the two options were chosen equally by participants, we can compute the chi-square value for each scenario. Table 3.1 shows these results for all 6 scenarios.

Table 3.1: Chi-square values for each scenario comparison, where O_L represents number of votes for the live recording, O_F the number of votes for the falsified recording, and χ^2_{cv} the critical value for two-tailed, one degree of freedom chi-square test (3.841).

Scenario	O_L	O_F	E	χ^2	χ^2_{cv} - χ^2
Female Voice (Chapel)	16	18	17	0.118	3.723
Male Voice (Chapel)	16	18	17	0.118	3.723
Orchestral (Chapel)	20	14	17	1.059	2.782
Female Voice (Meadow)	17	17	17	0.000	3.841
Male Voice (Meadow)	22	12	17	2.941	0.900
Orchestral (Meadow)	16	18	17	0.118	3.723

χ^2 is compared to the critical value for a two-tailed, one degree of freedom chi-square test, which is equal to 3.841. For each scenario, the resulting value is positive. This means that even in the case of the Male Voice (Meadow) scenario, the votes for the live and falsified recordings are still within a reasonable range of each other. This supports the hypothesis that real-world audio experiences can be recreated to such a degree that VR users struggle to distinguish live from falsified performances.

Furthermore, spectrograms were generated to show visual comparisons between the live and falsified audio scenarios. Figure 3.9 compares the live and falsified orchestral Chapel recordings as an example. The general shapes of both graphs are very similar. The lower pitched starting melody appears almost identical between the two recordings. At around 20 seconds in, a higher melody begins, and even though the falsified recording appears to be slightly less defined, with more higher frequency signals coming through, at lower frequencies, the recordings show comparable signals. This supports the obtained results, as there was only a 3-vote swing between participants choosing between the real and falsified orchestral Chapel scenarios.

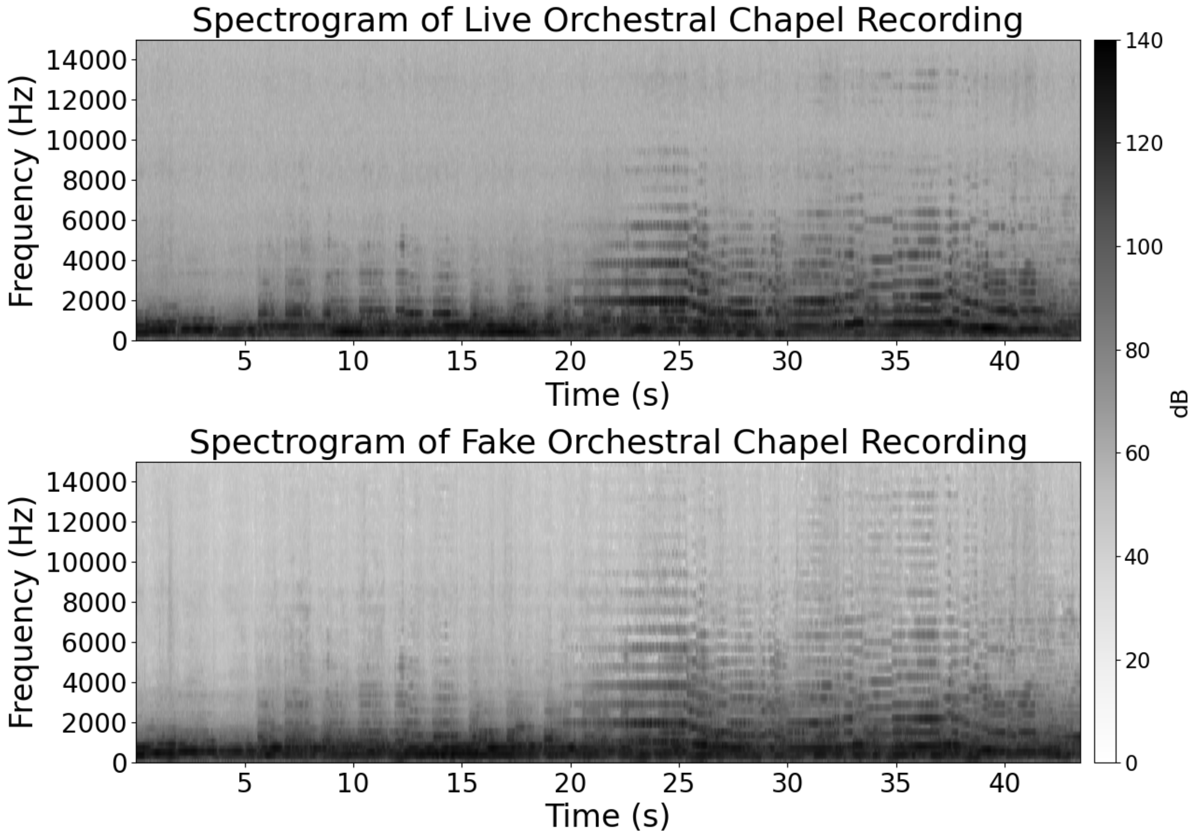


Figure 3.9: A pair of spectrograms comparing the live and falsified orchestral Chapel recordings.

There were many practical limitations when it came to the running of this study. The task of recruiting participants was non-trivial as it was reliant on the participants themselves to have a VR setup. Although virtual reality has become more mainstream in recent years, this will have limited who could take part in the study. The participants will have mainly consisted of people who use Reddit frequently and own a VR headset, therefore being familiar with virtual reality. This also meant that a variety of different headsets were used, with the most popular being the HTC Vive, so headset consistency and headphone quality could have influenced the results. However, there did not appear to be any strong correlation between the type of headset used and the ability to determine the live audio recording. Table 3.2 shows that for all seven headsets used, participants on average correctly guessed the live recording for three or four of the six scenarios.

Table 3.2: Table comparing type of headset to average correct identifications of live audio recording

Headset Used	Number of Participants	Average number of times (out of the 6 scenarios) the participant identified the live audio (nearest 0.5)
HTC Vive	16	3
Oculus Go	6	3
Valve Index	5	3.5
HP Reverb	2	3
Pimax 5K	2	3.5
Oculus Rift S	2	3.5
Oculus Quest	1	4

As is expected, there are still practical limitations of VR setups which prevent fully immersive experiences from being achieved, as only visual and auditory senses are engaged. This means that for the meadow scenario, there is a missing factor of the physical feeling of wind, or air temperature, to list a couple of examples. Furthermore, for practicality, these scenarios were only recorded at one height relative to the ground, so taller or shorter participants may have felt less immersed in either experience. Finally, recording equipment such as the microphone and tripod can be seen if the participant looks down as opposed to a virtual representation of themselves. None of the 34 participants mention this at all within the feedback surrounding immersion, however, it is a factor that could limit how authentically a real space is recreated virtually. A cube-shaped camera array could help to improve the experience as the downwards facing camera's recordings could be replaced by a still image or video where the tripod is not visible.

3.6.1 Further exploration and improvements

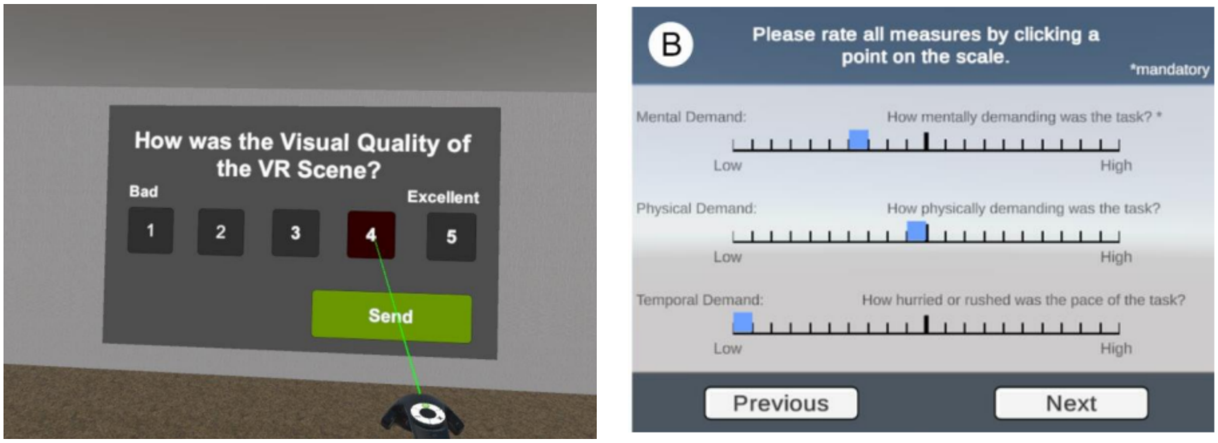
One of the core improvements that would be beneficial to make for this study would be to wrap the whole experience into a single program that does not require the participant to remove the headset during the tasks. Ripping subjects out of VR after every pair of videos breaks the immersivity they may have built and adds extra time to completing the study itself, as they are constantly having to take off the headset, answer some questions, and re-enter VR multiple times.

To improve upon this, utilising Unity and turning the experiment into a single file experience with a built-in form to fill in would have been preferable in many ways. At the time of creating this study (2019) there was no in-built method of utilising Unity's software with a 360° video with ambisonic audio. Even as of autumn 2022, there is still no in-built spatial audio functionality [164]. However, Resonance Audio can be used alongside Unity's Skybox component to create a spatialsed 360° visual/audio experience [165].

When it comes to integrating a questionnaire within Unity, this is once again not possible within just Unity. However, two projects have emerged that used Unity3D to create a questionnaire asset that can be integrated with VR environments.

VRate, now under the umbrella of VRTK, was created to combat the issue of assessment in VR scenarios being cumbersome and breaking immersion when a participant is forced to repeatedly take off a headset to respond to questions [8]. Similarly, *VRQuestionnaireToolkit* is a second VR questionnaire toolkit created in Unity3D, however, this open-source tool comes with some pre-installed standard questionnaires that can allow users to give more in-depth and varied responses to questions posed [9]. Examples of the two different user interfaces can be seen in Figure 3.10.

Furthermore, using Unity over Vive Cinema would allow for a more personal experience in some ways when engaging with the virtual environment. By using a game engine such



How was the Visual Quality of the VR Scene?

Bad 1 2 3 4 5 Excellent

B

Please rate all measures by clicking a point on the scale.

*mandatory

Mental Demand: How mentally demanding was the task? *

Low High

Physical Demand: How physically demanding was the task?

Low High

Temporal Demand: How hurried or rushed was the pace of the task?

Low High

(a) (b)

Figure 3.10: (a) VRate's UI for adding a questionnaire into Unity [8]. (b) VRQuestionnaire-Toolkit's UI for Unity using NASA TLX [9].

as Unity, users can be given a body, which allows for height to be based on the person participating in the experience as opposed to being set to one height.

This concept does not interact well with a skybox/skysphere however, as these options treat the image surrounding the user as being infinitely far away from them. Therefore, this would not work with a physical body to represent the watcher. A potential solution would be to utilise small, head-height spheres to try to imitate this spherical scene more naturally so that it feels like you are virtually present in a physical space as opposed to just watching a video that is infinitely far away. Figure 3.11 shows a bird's eye view of this potential idea being used with 360° photographs in Unreal Engine, that could be adapted to play video.

The concept of making one program to fit all would rule out the use of headsets that do not engage with Steam VR, such as the Oculus Go, unless a researcher was able to export the program separately to be compatible with each headset's requirements: a time consuming and potentially impossible task for some headsets. However, in an environment where in-person studies are viable, this would not be a problem, as only one device could then be

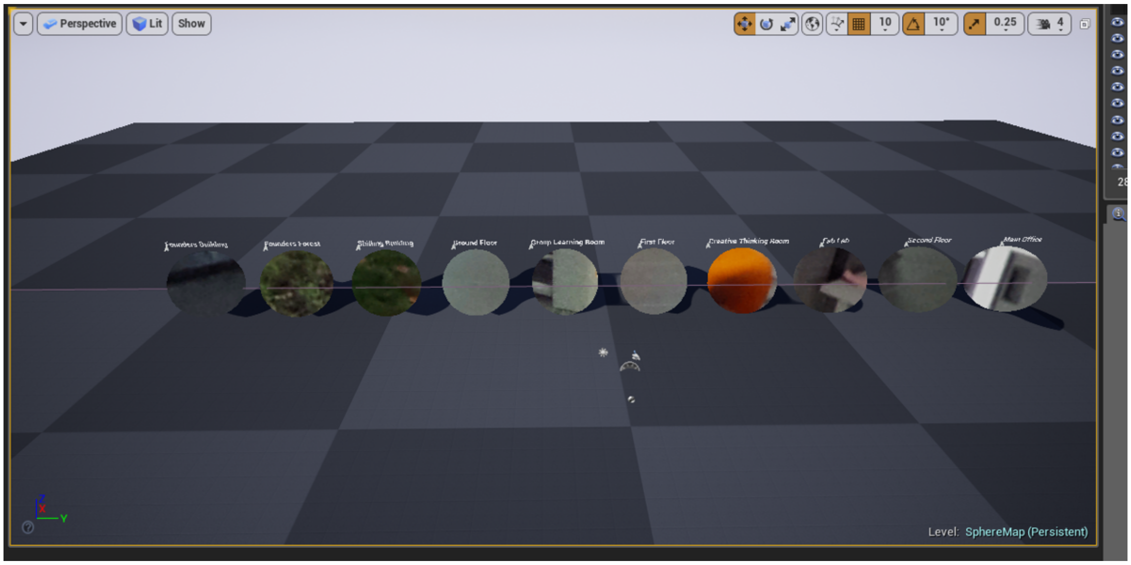


Figure 3.11: A scene created in Unreal Engine demonstrating the concept of utilising head-high spheres for a virtual body to walk into to combat the feeling of a Skysphere feeling infinitely far away from the user.

utilised for multiple participants to engage with, unlike how this study was forced to rely on other people having access to a headset of any variety.

Many participants stated in their responses that specific audio factors or cues they picked out from the recordings influenced their decisions when picking which recording live. Some of these statements do not actually align with what was actually present. For example, one participant stated that they chose a specific recording for the female meadow voice scenario as they could hear bird song in one and not in the other, despite both containing audible birdsong. Originally, there was a desire to randomise the order in which participants watched the videos for each pair, labelling them with non-order specific titles, such as non-number or letter characters, and varying the order in which they interacted with them. This was to mitigate potential unconscious bias when picking responses. With the study being altered to run without a researcher present, this was not viable to do without personally contacting each volunteer with their own instructions, increasing the barrier of entry for taking part.

On reflection, instead of randomising the order in which participants watched videos, it could have been advantageous to give different participants different subsets of videos. In particular, not always giving a participant the live and falsified recordings to compare, but the live or falsified one twice to see if participants still rationalise their decisions based on differences they can hear, even when there are not any. This would require a different set of questions to be posed to the participant but would not require major alterations to the study. This method would have been harder to implement at the time where a researcher could not always be present, and a larger sample size of participants would have been more appropriate to counter the three variants each scenario now could have.

3.7 Summary and conclusions

By measuring the impulse response of a physical space, the audio cues of that space have been recreated to such a degree that it is difficult to tell live from falsified performances. The system implemented is capable of simulating both indoor and outdoor physical spaces so that regular VR users struggled in both cases to distinguish the live from the falsified responses. The method demonstrated employed commercially available equipment and software, utilising REAPER, Adobe Premiere Pro, and Vive Cinema to create the final project.

Section 3.1 explained the core goals of the study, alongside the motivation behind the purpose of this research. The next section detailed how all of the audio and video recordings were collected, from the creation and recording of the impulse responses required to imitate locations aurally to the dry recordings needed for consistency throughout the study. In Section 3.3, the video and audio rendering processes were outlined, explaining the nuances of multiple pieces of software that are able to interact with ambisonic audio and 360° video, such as Facebook 360 Spatial Workstation and Premiere Pro.

The procedure behind the study was then presented, exploring the set-up and method behind the study and what would be expected of participants. Differences between the initial pilot study and the final experiment were explained. In Section 3.5, the results of the study were displayed and analysed, confirming the original hypothesis. Discussion on potential reasons behind responses and irregularities were discussed in Section 3.6, alongside further attestation behind the methods utilised for the creation of this experiment. Finally, further research ideas based on these results and improvements to the initial study were outlined.

Chapter 4

Measuring and Modelling Physical Spaces Virtually

> *"Sound and space are inextricably connected, interlocked in a dynamic through which each performs the other, bringing aurality into spatiality and space into aural definition."*
>
> - Brandon LaBelle, *Background Noise: Perspectives on Sound Art [166]*

Chapter 3 introduced the concept of using simple acoustic measuring techniques to recreate sounds recorded in a physical space virtually. The chapter revolved around stationary experiences, where participants were required to listen for differences and not physically move around a virtual space. As described in Chapter 2, there are programs such as Google Resonance Audio SDK (Resonance Audio) that can calculate the IR for a space through its geometry and materials in virtual reality, and on the fly if required. These tools are used often in video games to make scenes feel more realistic with directional and natural audio within the world, for example by adding bird song to a moving bird object, or the sound of trees rustling in a forest. The developers of Resonance Audio encourage recording real-world sounds to integrate into projects and provide an example scene with acoustically varying spaces such as a bathroom and a cave [167]. The core sound spatialisation algorithms used in the program are detailed in an accompanying paper [31], and Resonance Audio has been used in other scientific studies [168–170]. However, there has not been any dedicated exploration of how well these programs recreate real-world physical spaces audially. Are they exact or just authentic?

Hence, a new study was devised: how good are programs like Resonance Audio at recreating acoustic models of reverberant spaces when directly compared to real-world data creating these models? The concept of telling a program the materials a space is made of and letting it predict how a space would sound could have implications on the architectural design of buildings, letting clients physically explore a potential new build aurally, if the program does indeed create a realistic audio model for a space. What are its limits however? Is the virtual environment comparable to the real-world equivalent?

Over the course of this PhD, it has become even more apparent how a person's perception of a situation is just as important as the physics behind the acoustics of a space. As was explored in Chapter 3 in regards to the Chapel Scenarios, some participants argued that the more reverberant sounding recording made them believe that that model had to be real as it was what they expected to hear, whereas others decided that it being "more echoey" meant it had to be the falsified one. This has also been corroborated by prior work in the field.

Experiences can be more "realistic", can be aiming to be as true to the real-world equivalent as they can, but still fall short when compared to sequences that aim to create an "authentic" experience. This "reality paradox" can be found in discussions around the creation of realistic video game experiences, and other media [34] [171]. In the Battlefield series of games, audio characteristics from what other media has caused a person to assume should be present have been appropriated in-game alongside a rigorous real-world audio experience modelling approach. Inauthentic audio characteristics were shown to enhance immersion while playing the games when they were used carefully. For example, the sound of the player's footsteps as they move around in-game are always present and are not audially blocked out by other louder audio sources like explosions, causing the player to feel a greater sense of perceptual presence within the environment. This ego-ludic (heard only by the player) sound assists in aligning what a player expects to hear, compared to what they would hear in real life.

The paper goes on to discuss the use of "emblematic sound(s)" within media that embody a certain environment, even if they are not an audio element of that space in the real world, and how the quality of the medium itself also informs our concepts of the authentic [172]. That authenticity is also rooted in how a person believes a sound would have been recorded and stored in a particular era or context [34]. In the case of the Battlefield games, the more "authentic" sounding audio came from the cheaper recording set-ups as they felt more believable to players. Therefore, it is unsurprising that, in the case of this thesis, when users were presented with a reverberant sounding recording alongside a less reverberant sounding recording, each person projected their own decision on which one was more immersive to them, based on what they decided was "authentic".

From the above groundwork, a second discussion could arise; would a virtual environment with conflicting visual and audio cues make it easier or harder for a participant to correctly identify the real-world audio model? An example of this would be if the participant was presented with the visuals of a stereotypically reverberant space with the accompanying audio models stemming from a less reverberant environment.

The chapter begins with a brief discussion about the goals and motivations behind the study and what was hoped to be gained from it. Section 4.2 discusses the creation of the model chapel physically, with Section 4.3 exploring the implementation and fabrication of the acoustics of the chapel. The method behind the experiment is reviewed in detail in Section 4.4, detailing the layout of the virtual environments and study procedures. The results are then discussed and analysed, followed by a brief overview of the research conducted and potential further research. The chapter then concludes with a summary of the hypothesis.

4.1 Sound produced from acoustic modelling cannot be correctly distinguished from sound produced from real-world measurements in reverberant virtual spaces

The goal of this study was to compare the perceived validity of sound fabricated using acoustic modelling techniques and sound produced using real-world measurements in varying virtual spaces. Participants would be required to change between two acoustic models in two visually different virtual scenarios and decide which acoustic model they thought was created from real-world data, and which of the two they thought sounded the most realistic for the reverberant space, whatever the visual surroundings looked like.

In each scenario, the participant could spend as long as they liked exploring the virtual space by playing two different sound sources: a female voice and a 4 person choral choir. The acoustic models for the space could also be swapped between while these sound sources were being played, and the participant is also able to freely move around the virtual space within a set boundary while this is happening. Like in previous chapters, due to the ease of access and information, especially over the lockdown, the Chapel at Royal Holloway University was selected for this study as it is a vibrant, resonant space that is often utilised for its impressive acoustics.

There are substantial precedents for measuring the acoustics of performance spaces to recreate them in VR or otherwise. Whether it is for making cultural heritage more accessible [173] [174], or exploring the acoustics of a space for varying sound mediums, such as speaking or singing [175–177], acoustics of spaces and their behaviours have been interpreted and applied to virtual experiences. These methods provide historians, musicologists, and others with new perspectives of these spaces, especially through the rise of VR software which has led to multiple acoustic digital reconstructions with three degrees of freedom, with six-DOF systems under experimentation as well. These auralisations are a snapshot in time. They are a static representation of how an environment sounds or sounded [174]. Unlike these examples, the Chapel is fully intact and maintained to be almost as it was when designed initially, giving us a way to compare the acoustics of a real, used space with both measured and fabricated acoustic models, as opposed to relics of past spaces that were designed to be acoustically vibrant but are not in their full condition.

4.2 Creating the visual chapel model

The Founder's Building at Royal Holloway, which hosts the Chapel inside, was built for Thomas Holloway and his wife using the proceeds of his philanthropy as a college for women. Opening in 1886, the architecture of the building was inspired by the Château de Chambord and other such châteaux in the Loire valley. Despite the founder insisting that the college be secular, a rather lavish, non-denominational chapel existed from the beginning within the main building [178]. The Choir of Royal Holloway was formed when the college opened, and daily sung morning services continue to run to this day. The Chapel has been the home of these services since the beginning, so it is no surprise that the acoustic design is as lavish as the gilded artwork [179].

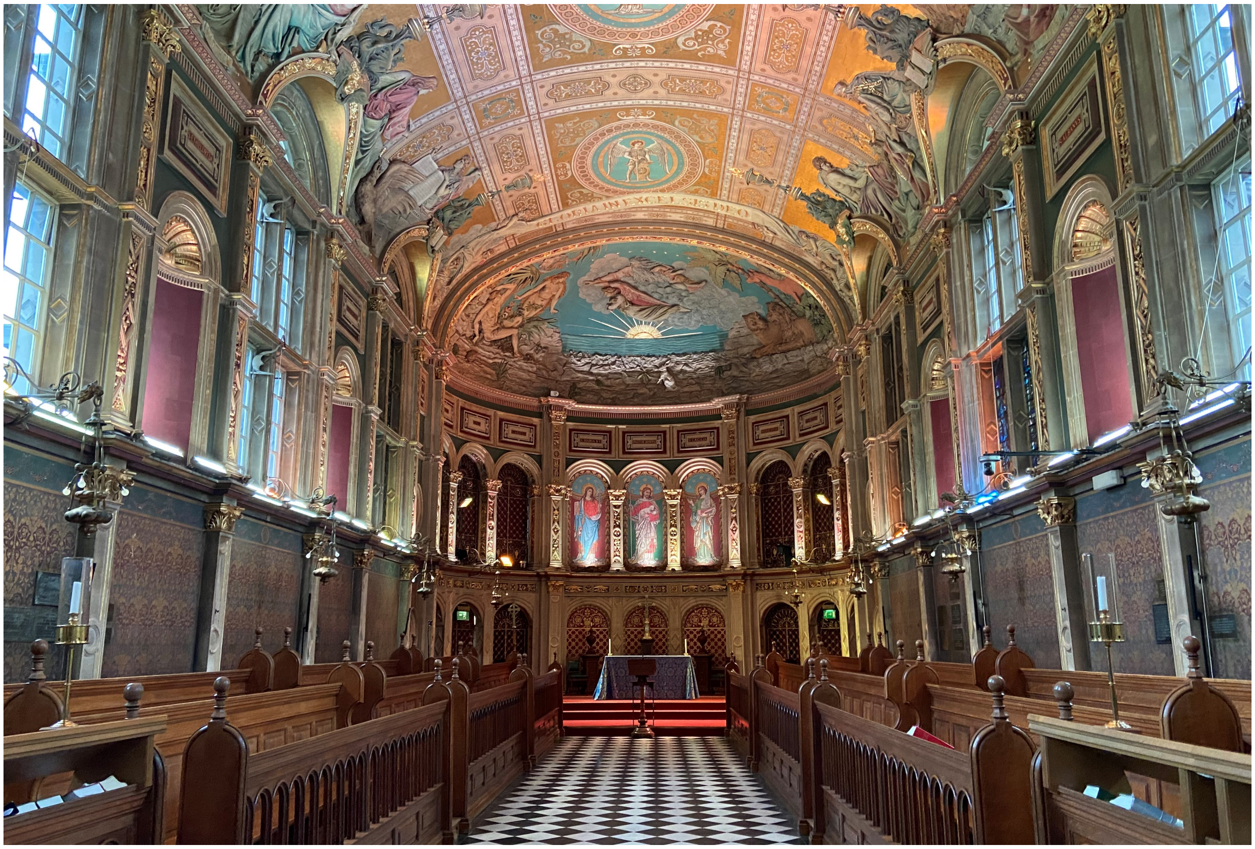


Figure 4.1: A photo of the Royal Holloway Chapel, 2021.

The Chapel is a magnificent space, and its magnitude must come across if created in a virtual environment if the intention is to immerse someone fully in that space. For this study, the aim was for the virtual Chapel to be able to be physically explored by participants, so it was important that the space resembled as closely as possible that of the real physical environment, which can be seen in Figure 4.1. The model that was created is dimensionally a one-for-one replica of Royal Holloway's chapel. The artwork and finer gilding was not included due to time restraints, but the grandeur of the space does come across as intended, especially when audio stimuli are incorporated.

4.2.1 CAD and modelling

Initially, internal dimensional information was not able to be gathered physically at the location. Figures 4.2 and 4.4 were provided by the CAD Technical Support team at Royal Holloway, along with some of the original sketches for the building (as seen in Figure 4.3).

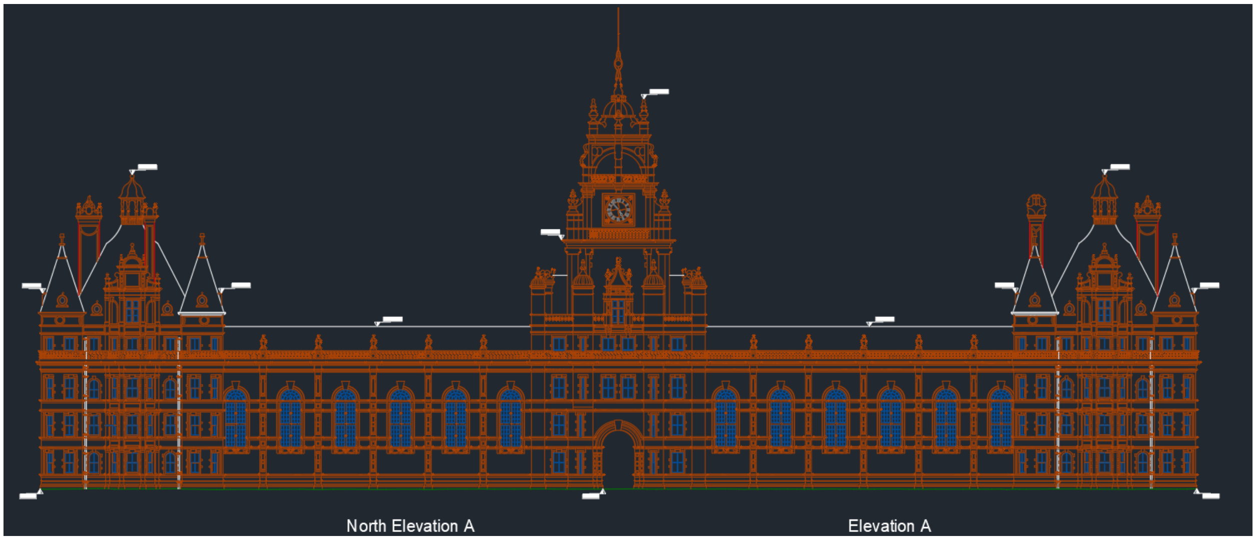


Figure 4.2: A dimensional drawing of the north side of Royal Holloway University of London's Founder's Building provided by the CAD Technical Support team.

Figure 4.2 helped to provide the size of the windows and their height above the floor inside the building. Real-world measurements of the pillars on the outer face of the chapel were measured to compare to this drawing to make sure that the dimensions were correct when scaled. But not all parts of the chapel could be measured in this way. The radius of the curved roof along with the domed northern corner were not attainable from the elevation drawing or CAD file, but thankfully, a side-by-side comparison with one of the original architectural drawings allowed for the radius to be found. This can be seen in Figure 4.3.

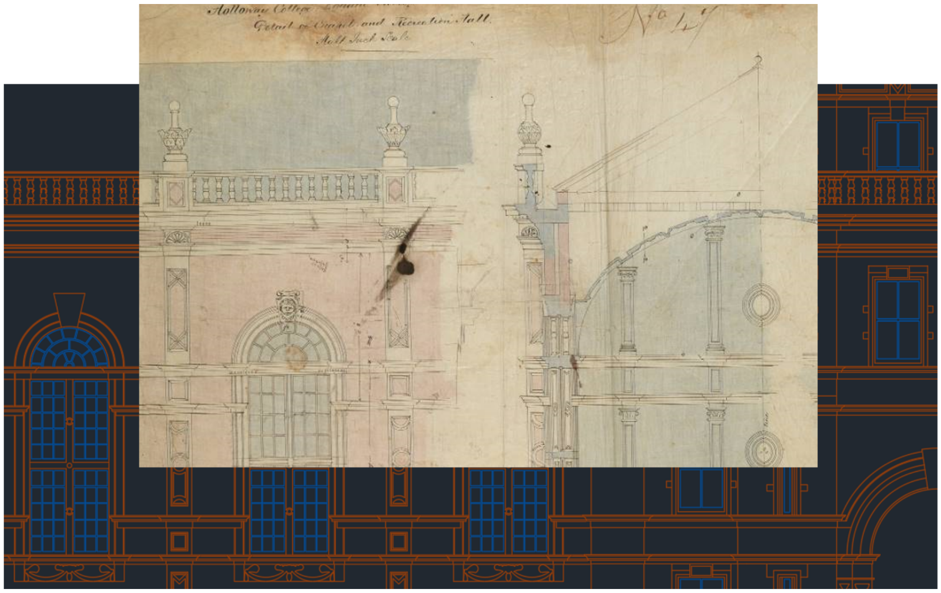


Figure 4.3: A side by side comparison of a CAD drawing of the chapel with one of the original architectural drawings. This was utilised to find the radius of the arched ceiling.

Floor plans of the chapel were also obtained (see Figure 4.4) and used to build parts of the internal dimensions of the space, such as the pillars beside each alcove and window. Physical evaluations at the chapel were made to measure internal objects and dimensions not present in either dimensional drawing, such as the internal pillar heights, step heights and widths, and dimensions of the pews.

Initially, an attempt was made to build the model in ProBuilder - Unity's structure building program, as it allowed for easy transfer into a VR scene. For the simple geometries, this worked fine, however for curves with varying radii, and for interlocking structures, the program struggled. Some of the features, such as the Boolean Tool allowing for multiple objects to either be joined, subtracted, or intersected, were experimental, and ultimately caused more problems than they solved due to some of the more complicated geometries

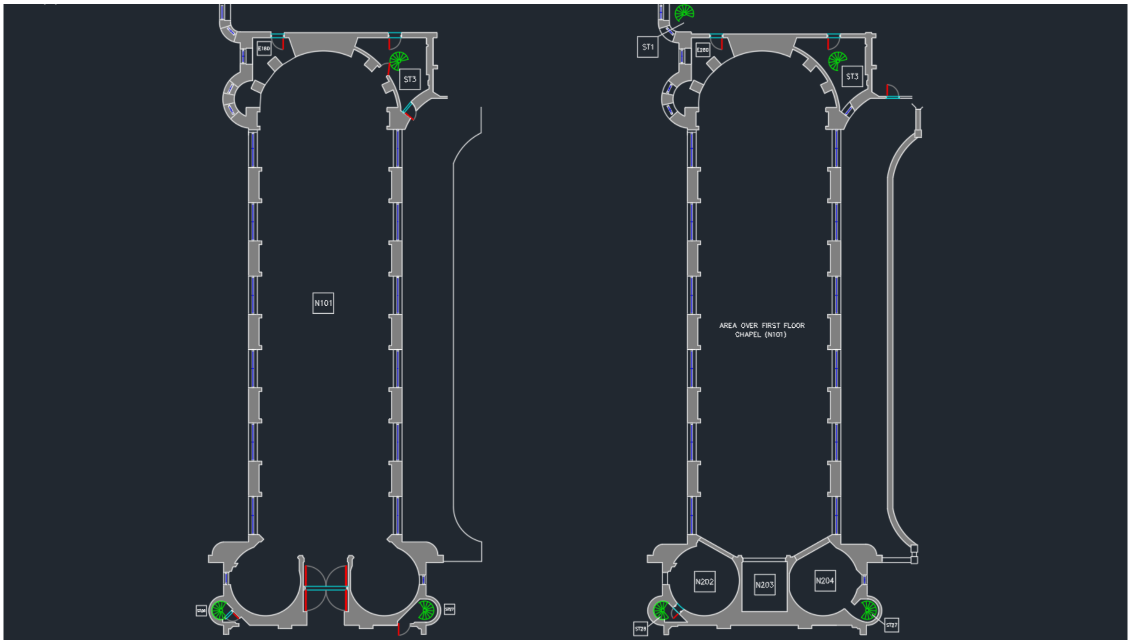


Figure 4.4: The floor plan for the ground and first floor of Royal Holloway's chapel.

involved. This prevented some objects from being manipulated past a certain point at which Unity decided that the object was "too complicated", which broadly meant that a 3D object was comprised of too many smaller faces. This was especially an issue where the roof and the window arches intercepted. Ultimately, ProBuilder was not suited to this specific task (see Figure 4.5). For less complicated structures, it would have worked.

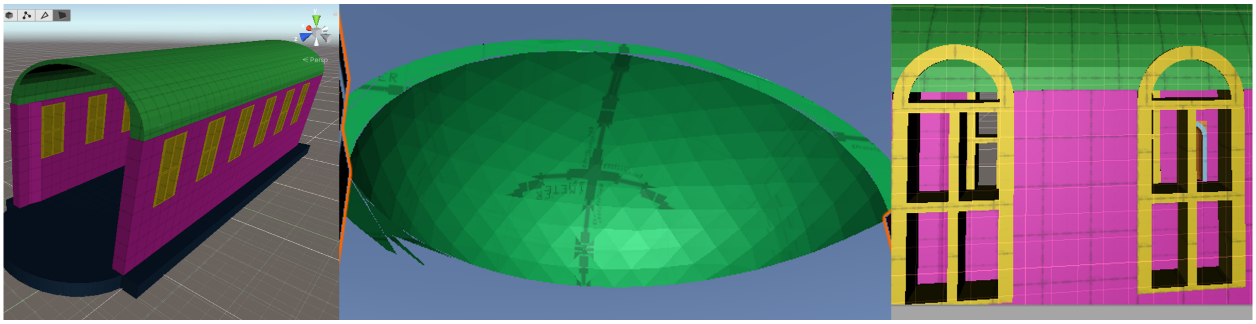


Figure 4.5: Initial chapel modelling in ProBuilder. The different colours are just for easy identification of separate components of the model.

The model was then rebuilt to completion in Autodesk Inventor Professional. This software provides professional-grade 3D mechanical design for objects of any size. Figure 4.6 shows the exterior and interior of parts of the raw chapel model. As it is only intended to be viewed from the inside, no attention was given to embellishing the exterior. The model itself is built to scale.

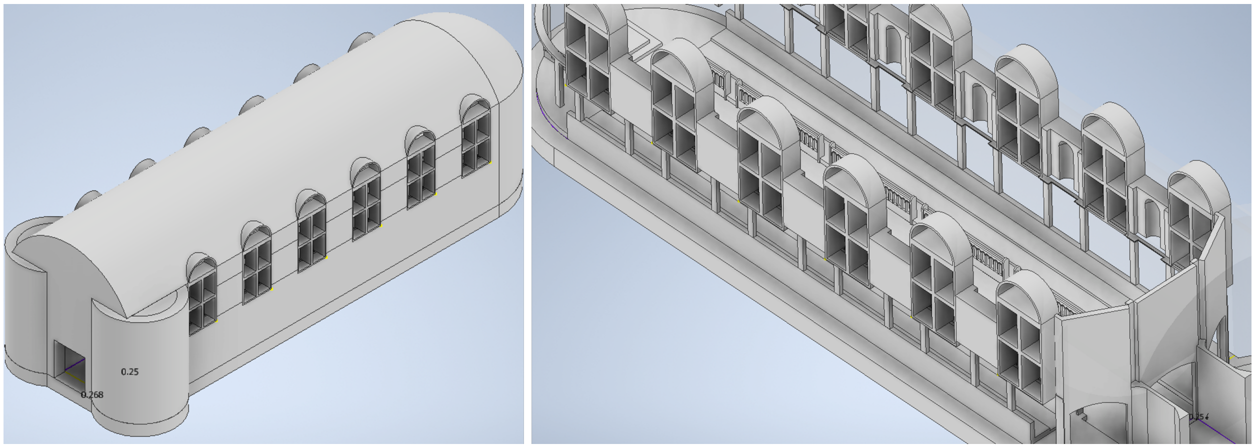


Figure 4.6: The model of Royal Holloway's Chapel made in Autodesk Inventor Professional. The figure on the right displays some of the detailing inside, which is where participants will virtually be while taking part in the study.

The Chapel is a large, reverberant space, so in general, only the major geometries for the space were built. This is because participants would be virtually locked within only the central body of the chapel. This means that they would not be able to get near to either the walls, the far edges of the room, or the small side rooms. Hence, some details, such as the furniture in the side rooms, were deemed inconsequential to this study and were left out.

4.3 Measurement and fabrication of RT60s

As described earlier in this thesis, there are multiple methods with which the reverberation time of a space can be measured. Due to limiting factors, such as the pandemic, and off of the success of utilising simple measurement methods in Chapter 3, the technique of popping balloons was applied.

As will be discussed in more detail below, Resonance Audio can create and measure the reverberation time for a range of frequency bands for a virtual space, and this technology has been discussed and employed in academia over the last few years; [31], [180], [181], to cite a few. Through researching how these fabricated RT60s are implemented, it was then assumed that these values could be changed manually so that real-world measured values could be added alongside the computer generated versions created by Resonance Audio. The program was also used for the 3D audio rendering due to its high-quality handling of both low and high order ambisonics [31].

4.3.1 Measuring Impulse Response frequency bands

To create and measure the IR for the chapel, like in Chapter 3, multiple balloons were popped and recorded. Unlike in Chapter 3, for this study, participants would be required to move around in the virtual space and not just be stationary. This meant that consideration needed to be made for where in the virtual chapel people would be allowed to explore. The ends and edges of the chapel, for example, have different acoustics from the main body of the chapel.

Three balloons were popped for each of three recording locations along the main body of the chapel in unoccupied conditions (except for a balloon popper and myself). This was done so that average frequency bands for the reverberation time could be ascertained.

The recordings were imported into Audacity along with Aurora Tools for Audacity; a set of plugins created by Angelo Farina that were initially developed for measuring and manipulating room acoustical impulse responses, performing analysis of acoustical parameters, and auralisation [182]. Using the plugin `Acoustical Parameters Calculator (ISO 3382)`, the measured results of the reverberation times and Early Decay Time (EDT) were processed and extracted. The results for each location were averaged, and when compared, demonstrated that the reverberation time in the middle section of the chapel is relatively consistent. These averaged values can be seen in Table 4.1 with any outliers indicated.

Table 4.1: Averaged measured results of Reverberation Time as defined by the standard ISO 3382-1. Anomalous values are indicated through red text.

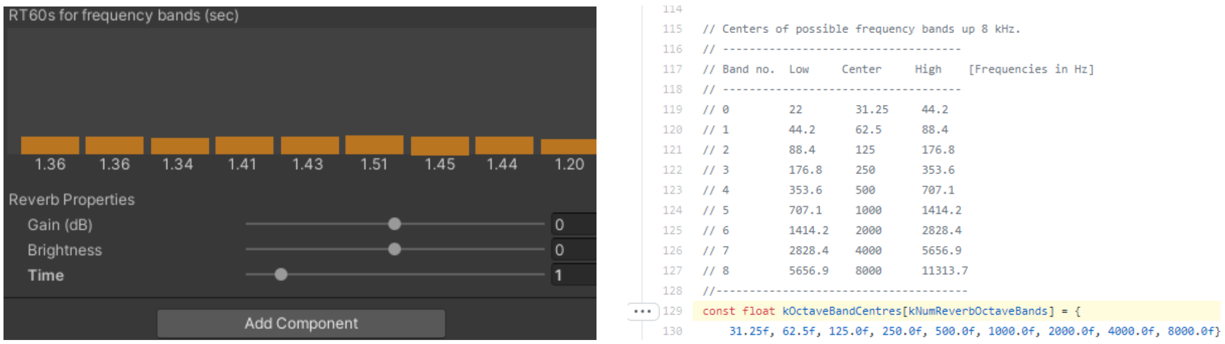
	Frequency Band (Hz)								
Location of Mic	31.5	63	125	250	500	1k	2k	4k	8k
1	0.861	1.032	1.656	2.153	2.344	2.373	2.256	1.853	1.292
2	0.021	1.715	1.828	2.296	2.497	2.555	2.282	1.977	1.441
3	0.817	1.706	1.902	2.435	2.474	2.706	2.261	1.939	1.301

4.3.2 Modelling Impulse Response frequency bands in Resonance Audio

The chapel model was exported as an OBJ file, which allowed for direct upload into Unity. Once uploaded, Resonance Audio was used to assign material properties to each component of the chapel [167]. For example, the pews were assigned as “wooden”, the floor as “tiled” etc. If the exact material was not present, the most similar material was chosen within the program. These were then used to determine the acoustic properties of the objects with Resonance Audio.

A reverb probe was placed in the central area of the chapel and its region of application was set to be the same as for the live impulse response measurements. Reverb probes define a location where the reverb properties of a space are sampled/computed. These properties are then baked using ray-tracing to simulate the way sound waves interact with the surrounding environment. This means that the acoustic properties for a space are precomputed and preloaded into the virtual environment. When the virtual listener is present within the region, the baked reverb is applied to what they listen to [167].

The RT60s for the frequency bands (sec) for the baked reverb probe were calculated by Resonance Audio. The values of the frequency bands corresponding to each time in resonance audio were not immediately obvious however. The interface within Unity only lets you change reverb properties such as the gain and brightness. It presents the frequency bands and the corresponding RT60, but not the specific frequencies of the bands. There is no way to find these answers in the front end of Unity. These values were eventually found within the source code for reverb probes within Resonance Audio, and ranged from 31.5 Hz to 8 kHz [183]. Figure 4.7 shows the interface and source code side by side.



Gain (dB) 0
Brightness 0
Time 1

(a)

```
// Centers of possible frequency bands up 8 kHz.
// ------------------------------------
// Band no.  Low     Center     High    [Frequencies in Hz]
// ------------------------------------
// 0         22        31.25     44.2
// 1         44.2      62.5      88.4
// 2         88.4      125       176.8
// 3         176.8     250       353.6
// 4         353.6     500       707.1
// 5         707.1     1000      1414.2
// 6         1414.2    2000      2828.4
// 7         2828.4    4000      5656.9
// 8         5656.9    8000      11313.7
//-------------------------------------
const float kOctaveBandCentres[kNumReverbOctaveBands] = {
    31.25f, 62.5f, 125.0f, 250.0f, 500.0f, 1000.0f, 2000.0f, 4000.0f, 8000.0f}
```

(b)

Figure 4.7: (a) Unity *RT60s for frequency bands* panel. (b) Source code for the frequency bands and their ranges.

These fabricated RT60 results can be seen in Figure 4.8 where they are compared to the values of the averaged live measurements for the same region of the chapel. The live measurements have a significantly more varied decay time across the full range of frequency bands, ranging from 0.839 s to 2.545 s. For certain bands, the real decay times are half that of the resonance audio-generated ones, and for others they are double. The effect of this disagreement between the RT60s will be discussed further alongside the results of the study.

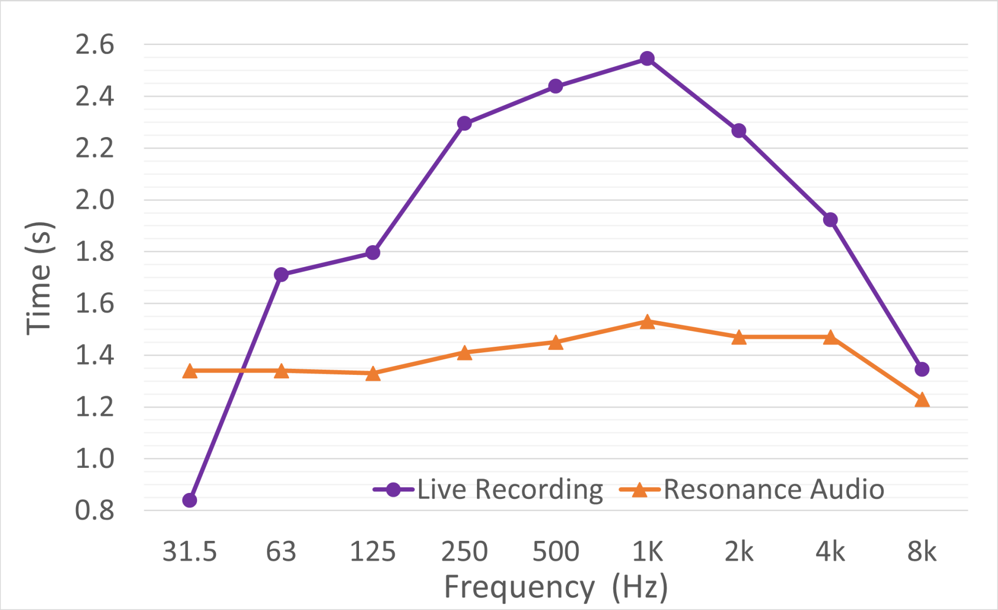


Figure 4.8: Measured (Live Recording) and Fabricated (Resonance Audio) results of Reverberation Time.

4.3.3 Implementation of the Room Impulse Responses

The RT60 values created by Resonance Audio were found to be created and modified from the file `ResonanceAudioReverbProbe.prefab` found in the Assets folder within the project's files. Within this file, these values are all stored as zeros until referenced within the code which creates the baked reverb. The RT60 values are then added to these when the probe calculations are done by Resonance Audio. It was discovered that as these values are just added, if the numbers are manually changed in the reverb probe file mentioned above, these are then applied to the probe, and therefore, change the reverberation time for the virtual space.

What this meant for the project was that by duplicating and renaming this file, we could create a pre-set RT60 for the frequency bands that could just be pulled in by the software and implemented in the region of application for a reverb probe. We could create two separate reverb probes, one each for the real and fabricated RT60s, that could be interchanged between, and updated, freely within the program, even while it was running. One probe running as normal, its values being created from the `ResonanceAudioReverbProbe.prefab` file, and the other being created from a new file named `EditableProbe.prefab`. The two reverberation models could be sampled on the fly, so a listener within the program could compare them by listening to, in this case, vocal samples.

4.3.4 Implementation of the sound sources

Unlike in Chapter 3, this study would allow participants to move freely within a certain area of the virtual space. This allowed for an exploration of not only various audio sources in a reverberant space, but also how they sound from up close and far away. To not overcrowd or overcomplicate the virtual play space, it was decided that only two types of sound would be implemented: a four-part choir and a short speech. The choir audio sources were placed in the pews to imitate a more natural performance for the space within the bounds of where the participant would be exploring. The speaking voice audio source was placed in the main body of the chapel in the aisle. Each set of sounds had a corresponding button with which to start the audio playing. These buttons were labelled and placed near to their respective audio counterpoints.

Even though the IR recordings were taken from the aisle, it was decided that the placements of the singers in the choir stalls would still be suitable for the experiment. The choir would be singing out towards the aisle and the participant play space, and not towards the wall or floor. The audio sources are placed near the tops of the singer objects so would emanate from above the pews. Participants would also only be listening to these singers from the aisle. Furthermore, as mathematical equations for estimating the RT60 for a space for a specific frequency band care about total room absorption for said frequency band, surface area of material in the space, and the volume of the space, it can be assumed that moving about one meter away from space will have a negligible effect on the RT60 as none of these factors will change [184, 185].

Based on the above observations, it was decided the initial RT60 measurement would be sufficient, so no separate measurements of the RT60s from the singer's locations were undertaken. Further measurements could be taken in future experiments to directly compare these RT60s.

The choir recordings were very kindly provided by Andrew Woodmansey, a Tenor singer who studied at Trinity College of Music, who had recorded himself singing a four-part arrangement of the song Calon Lân by John Hughes for an assessment. The voice source was reused from the first study and is a 15 second clip that describes the chapel. These sound sources were home recorded to be as acoustically neutral as possible, without access to a recording studio, so that they could have reverb added to them in the virtual environment. The global pandemic necessitated this less formal recording process for both this study and Andrew's assessments.

These sound sources were then attached to egg-shaped objects representing the singers and speaker. One of the big drives for using Resonance Audio is its ability to have directional audio sources. It was decided which way the singers and speaker should be facing, and this was matched in the software. Resonance Audio can simulate occlusion effects and directivity patterns, as is discussed in their literature [186]. The listener's location in relation to a sound source as well as the direction of the sound source itself changes how a sound sounds. If you walk around a person who is speaking, they will appear at their loudest when you are on the side that their mouth is, the side that the audio is being projected from. From behind, their body occludes some of this sound, and they are projecting away from the listener, which makes them sound different. Resonance Audio lets you change the directivity pattern for a source and mimic the ways in which a source would emit sound in the real world.

4.3.5 Discovered issues and solutions

Resonance Audio is still being updated and modified so despite all of its innovative capabilities, it is not always perfect. While creating this study, we ran into a couple of problems that are worth highlighting, as well as what was done to solve or work around them.

As more audio sources were added to the two scenarios A and B, many of these sounds stopped working properly. They would do one of two things: either not play at all when triggered by a button, despite Unity thinking that they should be playing; or the source would play and interact with the reverb probe, however it would stop being directional and just play in both ears in a mono configuration.

While researching into this problem, it was discovered that Resonance Audio claim that their digital signal processing algorithms are "optimized to spatialize hundreds of simultaneous 3D sound sources without compromising audio quality", however in practice, this does not seem to be the case [118]. From researching, this was not the first time that these issues had been encountered by people when using this program, and it seemed that there was a limit on how many audio sources you could add to a scene.

Once at the threshold, even if the audio was not actively playing, the presence of it alone would cause the other sound sources to stop functioning as intended. The developers claimed that the 2019.1 update should have fixed the issue, which it had not, and that for earlier versions of Unity, users should effectively deactivate the spatialisation for a sound when not playing it and reactivate prior to activating it with `.SetSpatialize(false)` and `.SetSpatialize(true)` respectively [187].

These methods were not successful within this project, and the discovered workaround for these issues was to completely remove an audio source from the scene when it was not playing. This was achieved through the buttons that triggered the audio to play.

When a sound button is pressed, it first stops all other audio sources from being played by removing the sound-playing objects from the scene. It was not possible to isolate just the audio part of an object to remove, therefore by doing this, the object does physically disappear. To counter this, there is a dummy object with no sound component always present and the objects with audio sources attached appear in the same location. This object-within-object does not affect the resonance audio reverberation calculations as the extra one is set as "invisible" to the sound calculations. This did fix the issue when running the program on a suitably powerful PC, for using both the HTC Vive and Index headsets. On a less powerful PC, the program had to be restarted whilst attempting the study as not all the sound sources would play. Restarting it did seem to fix the issue, but it seems that there is a core fault with Resonance Audio's optimisation.

The second bug discovered revolves around an audio source to teach users how buttons work. On loading the programme, participants enter a Hub World which leads to the two scenarios. This virtual room contains two example buttons to physically demonstrate how to interact with them throughout the study. These buttons were labelled "Start" and "Stop" and, once pressed, would play and pause an example audio source.

For some still unknown reason, when the program was initially loaded, the "Start" button did not cause the sound to play. If the "Stop" button was pressed first, then the "Start" button did work afterwards. These buttons were implemented differently from every other button in the program as instead of having one audio source attached to one button, the ability to stop the recording playing was added to a second button. All other sound sources are stopped by another audio source being activated. Despite numerous hours of searching, this bug has just become a feature, as nobody could work out a reason why this was the case.

4.4 Method

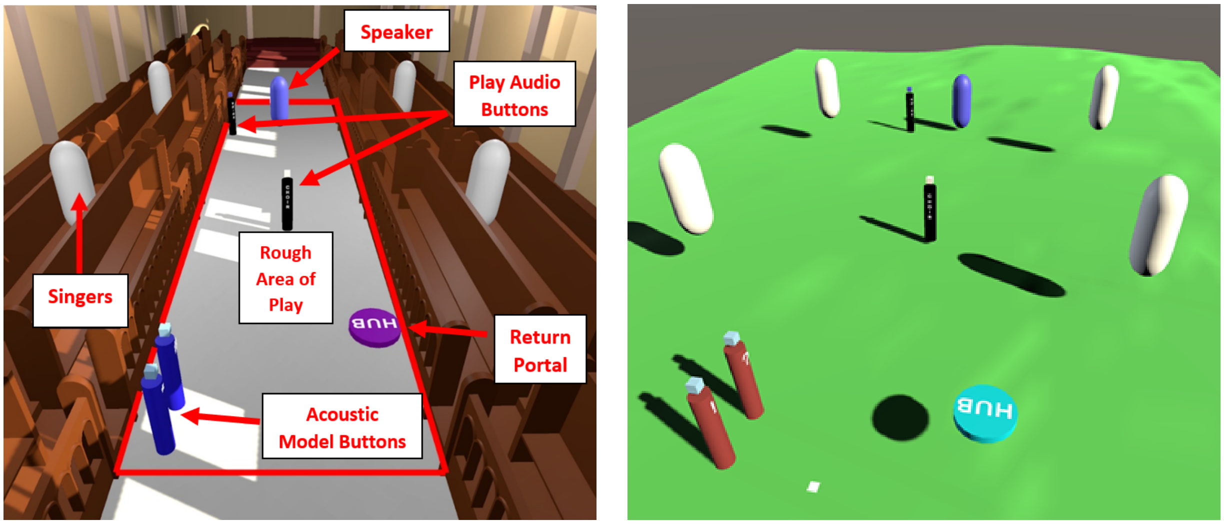
4.4.1 Layout of virtual space

Participants first enter a Hub World when they load the program. This hub world gives a few additional instructions on how to interact and move in VR and prompts the participant to go and read the instructions document if they have not done so already. There is an example button to press to demonstrate how to use future buttons in the scenarios, which also gives the user a chance to adjust their volume if they felt it necessary, as well as an explanation on how to teleport about. The level of playback is set to the same initial level by the researcher for each participant, which ended up being the maximum volume of the system. Once the participant is comfortable, they can enter Scenario A through a portal.

Scenario A is visually set in the chapel; the layout of which can be seen in Figure 4.9a. Participants are told in the instructions to press one of the two acoustic model buttons before proceeding to interact with the sound sources. These buttons are labelled with symbols as opposed to letters or numbers, and were positioned next to each other just to the left of where a participant arrives in the scene so as to not cause any bias towards deciding which one to press first. Participants are also unaware of which button correlates to which model.

There are no restrictions for when participants are allowed to change between audio sources and acoustic models, and in the instructions, they are encouraged to spend as long as they want comparing everything in each scenario. The singer and speaker objects (the egg-shaped objects) are directional sound sources, so are louder at the front than the back, and although it was not possible to get behind the choir singers, there was space to move fully around the speaker.

The position of the Choir and Voice buttons (which trigger the playing of the choir and voice audio files respectively) were placed in places which would provide a good starting point



(a) Scenario A

(b) Scenario B

Figure 4.9: Two figures showing from a birds-eye view, the two scenarios A and B.

for the listener to engage: the choir button is within the four singers, so participants would be able to hear the voices from all sides, but it was closer to two of them to hopefully encourage the listener to understand that, for example, the closer they get to a singer, the louder their voice becomes; for the speaker, the button was just to the side so that participants could easily go in front and behind the speaker.

Scenario B is visually very different and was made to give contrast between what participants were seeing compared to what they were hearing. The layout of the key components and interactions (the buttons, sound sources, IRs, the play space) were kept the same positionally, with the acoustic model buttons labelled with different symbols, * and # for Scenario A, and ! and ? for Scenario B. The rest of the map was just an open landscape, as can be seen in Figure 4.9b. Participants were intentionally not informed if the layout of the two acoustic models was kept the same (left and right buttons). The decision on which corresponded to each model was left to a coin toss for both scenarios to prevent any bias from the researcher.

4.4.2 Procedure

Participants were given an instructions document to read through that gave them an indication of what to expect once they entered the virtual space as well as what was expected of them.

They were told they would be presented with two scenarios: A and B and were given details pertaining to the two acoustic model buttons as well as the sound source buttons. They were also made aware that although the two scenarios would differ visually, for both cases they would be asked the same things: which acoustic model they thought was created from real-world data, and which of the two they thought sounded the most realistic for the reverberant space, whatever the visual surroundings looked like. As participants were not expected to have any prior VR experience, as well as visual depictions of what certain processes (i.e., pressing buttons) would look like, they were also given an interactive tutorial within the hub world once they loaded in. Participants were told that they must choose to activate one of the two acoustic models in the scenario before listening to any of the recordings. No data was collected about which buttons were pressed first in either scenario. They were then asked to complete the questions about Scenario A before moving on to Scenario B, and were told they could spend as long as they liked exploring each scenario. However, no one spent more than 8 minutes in either one.

All but two of the participants completed the study with the researcher physically present, so rather than having to take off the headset to answer the questions, they were read aloud, and the participants answers taken down verbatim. After the scenarios are explored, participants were invited to answer two further questions to do with the study, directly comparing the two scenarios' visuals and how that changed the way they perceived the audio environments. The participants typed their own thoughts for this section.

This study was also shared online for people who were already familiar with VR to partake in themselves. Their instruction document had an additional section illustrating potential bugs that may arise while taking part in the study, as well as avoidable issues that arose when during previous run-throughs of the study. This included things such as being able to glitch outside the intended play area quite easily and what to do if that happened, and the known bug mentioned in Section 4.3.4 with the Start and Stop buttons in the Hub World, to name a couple.

Because of this, like in Chapter 3, only a single iteration of the study was created, with all participants having access to the exact same copy of the study. This means that the orders of the scenarios and allocation of acoustic models to buttons were the same for everyone. As with the improvements detailed in Section 3.6.1 for the previous experiment, it would have been ideal to be able to randomise the order in which participants interacted with the scenarios and acoustic models to mitigate any potential bias across the answers. With the study being run without a researcher present in some cases, and with timing restrictions enforced on this research due to the global pandemic, multiple iterations of the study were not run. Measures, such as labelling the acoustic model buttons with symbols that did not lean towards a specific order, were taken to limit potential biases, however there is a chance that these results are not representative of the conditions.

4.5 Results

4.5.1 Participants

22 participants took part in the study, providing written informed consent, in accordance with the Royal Holloway Research Ethics Committee. Primarily students were encouraged to take part in the study, so many of them were not familiar with virtual reality prior to this study. Neither gender nor age were considered for this study as it was deemed inappropriate.

4.5.2 Questionnaire

Participants were asked a set of 5 questions for each scenario:

1. Of the two acoustic models which do you think was created with real-world data? (i.e. by physically taking measurements at the location) (*Required*)
2. How confident are you in your answer? (*Required*)
3. Why did you pick these answers? (feel free to leave blank)
4. Of the two acoustic models which do you think felt the most 'realistic' for a reverberant space? (this does not have to be the same as your previous answer) (*Required*)
5. Why? (*Required*)

And then a set of general questions inquiring about the effect that the visuals in the two scenarios had on influencing their answers:

1. Did the visuals in Scenario B make determining which acoustic model you thought was real easier, the same, or harder to ascertain? (*Required*)
2. Please explain your answer
3. Did the visuals in Scenario B make it easier, the same, or harder to pick the most 'realistic' acoustic model for a reverberant space? (*Required*)
4. Please explain your answer

Participants were also asked which VR device they were using to complete the study as well as a set of questions inquiring what their main area of study is/was and, if applicable, what sort of music they like to listen to. This was to see whether participants who listened to choral music more often, or were music or music technology students, tackled the study differently due to familiarity with the sound of music in reverberant, religious spaces.

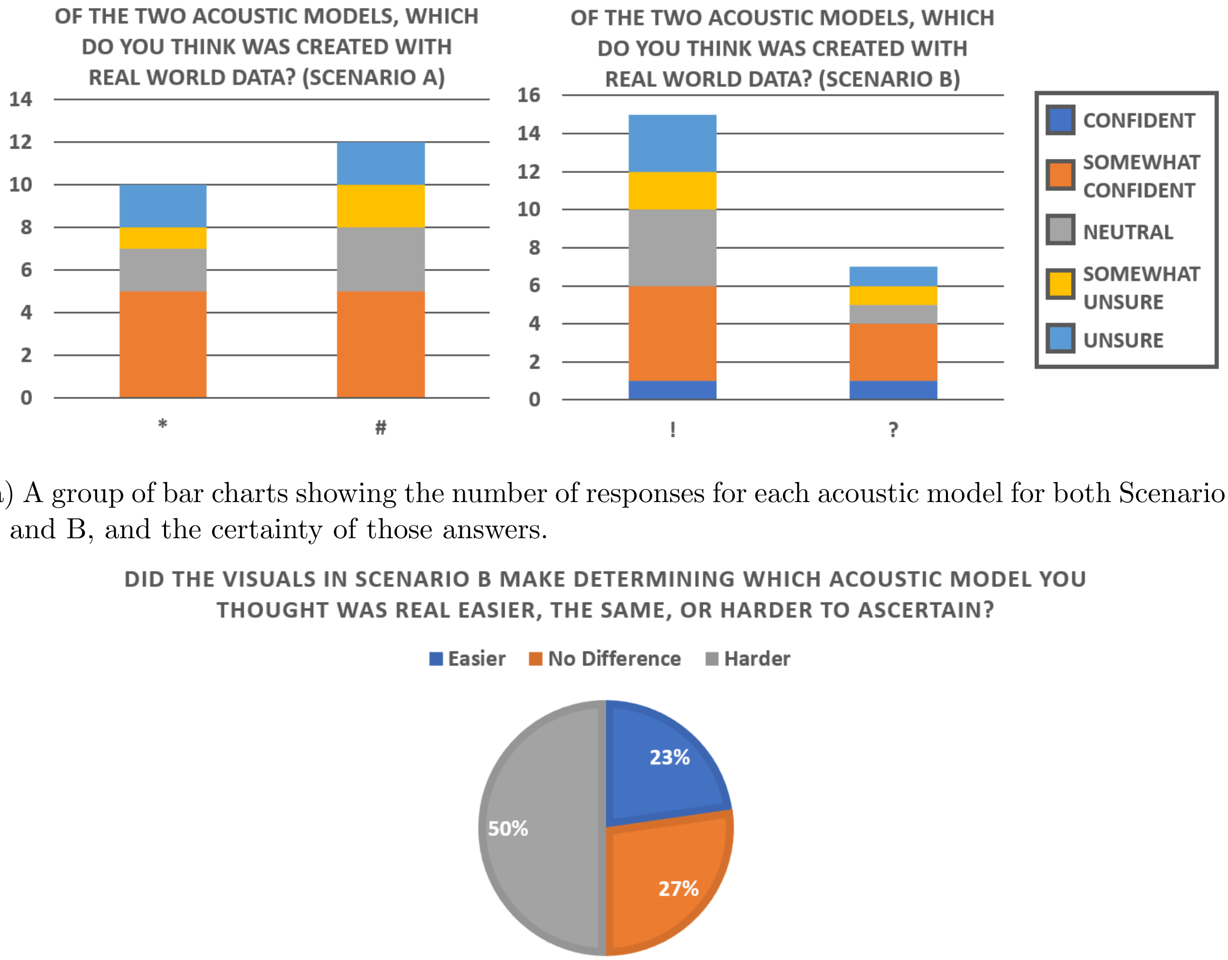
Participants were able to provide an answer of *Confident, Somewhat Confident, Neutral, Somewhat Unsure,* or *Unsure* when asked how confident they were in their answers, with elaboration left for other questions. Once participants had familiarised themselves with the study instructions, had the chance to ask any questions, and had filled in the consent form, they entered the virtual environment. Once comfortable with the VR headset, participants were given as much time as they needed to explore the Hub World, which gave them a chance to become familiar with moving around in a virtual space and how to interact with audio buttons. They were encouraged to not only move virtually by teleporting, but also by physically moving around. Most participants were students from the university, and many had never used VR before, so this was a useful exercise in becoming comfortable in a virtual space.

4.5.3 Questionnaire results

A total of 22 individuals took part in this experiment. Of these, only 2 took part remotely without a researcher present. Figure 4.10 shows the number of responses for the questions that asked about which acoustic model participants thought was created from real-world data, with the associated percentages for the bar charts given in Table 4.2 for clarity.

In Scenario A where the visuals that accompanied the audio were that of the Chapel, the results were almost a 50/50 split, with 45.5% of participants choosing the correct model, labelled *. In Scenario B, the difference in opinion was far less balanced, with only 31.8% of the responses being correct.

This is reflected in the results for Question 1 of the general questions, as shown in Figure 4.10(b). 50% of participants found that the change in visuals in Scenario B made it more



(a) A group of bar charts showing the number of responses for each acoustic model for both Scenario A and B, and the certainty of those answers.

(b) A pie chart showing the results regarding the difficulty for determining the answer for Scenario B for which model participants thought was made from real-world data.

Figure 4.10: The results for the questions about choosing which acoustic model was made from real-world measurements.

difficult to work out which acoustic model was the real-world one.

Table 4.2: Percentages of *Confident, Somewhat Confident, Neutral, Somewhat Unsure,* or *Unsure* responses for the questions on choosing which model was made from real-world data for both scenarios.

	Participant Certainty (%)				
Chosen Model	Confident	Somewhat Confident	Neutral	Somewhat Unsure	Unsure
*	0.0	22.7	9.1	4.5	9.1
#	0.0	22.7	13.6	9.1	9.1
!	4.5	22.7	18.2	9.1	13.6
?	4.5	13.6	4.5	4.5	4.5

For those individuals that chose * as the model that they thought was real, common comments related to greater "echoeyness" and a more "dynamic" or "harsher" sound to the model for the reasoning behind their choice, with a single participant describing the other model as "too full" comparatively to *.

The reasons provided for choosing the model labelled # however was far more varied. 25% of these responses were focused on why they thought * was fake. Comments such as "* was more clearer..., so felt more generated", and that "* had more reverb" were given as reasoning for choosing #. 50% of the other responses described # as more reverberant, more realistic, or more varied and detailed. There were comments from both sides, mostly from responses that were accompanied by a certainty of *unsure* that mentioned how the models were really similar, and that they "just chose", with one of these responders going on to say how, in their opinion, "...the difference between the two was negligible. You could have told me either were real and I would have believed you."

One of these IRs was mathematically more reverberant than the other (see Figure 4.8 for a reminder), yet many people described the less reverberant, generated acoustic model (#) as more full and echoey, when it just is not in comparison to the other model.

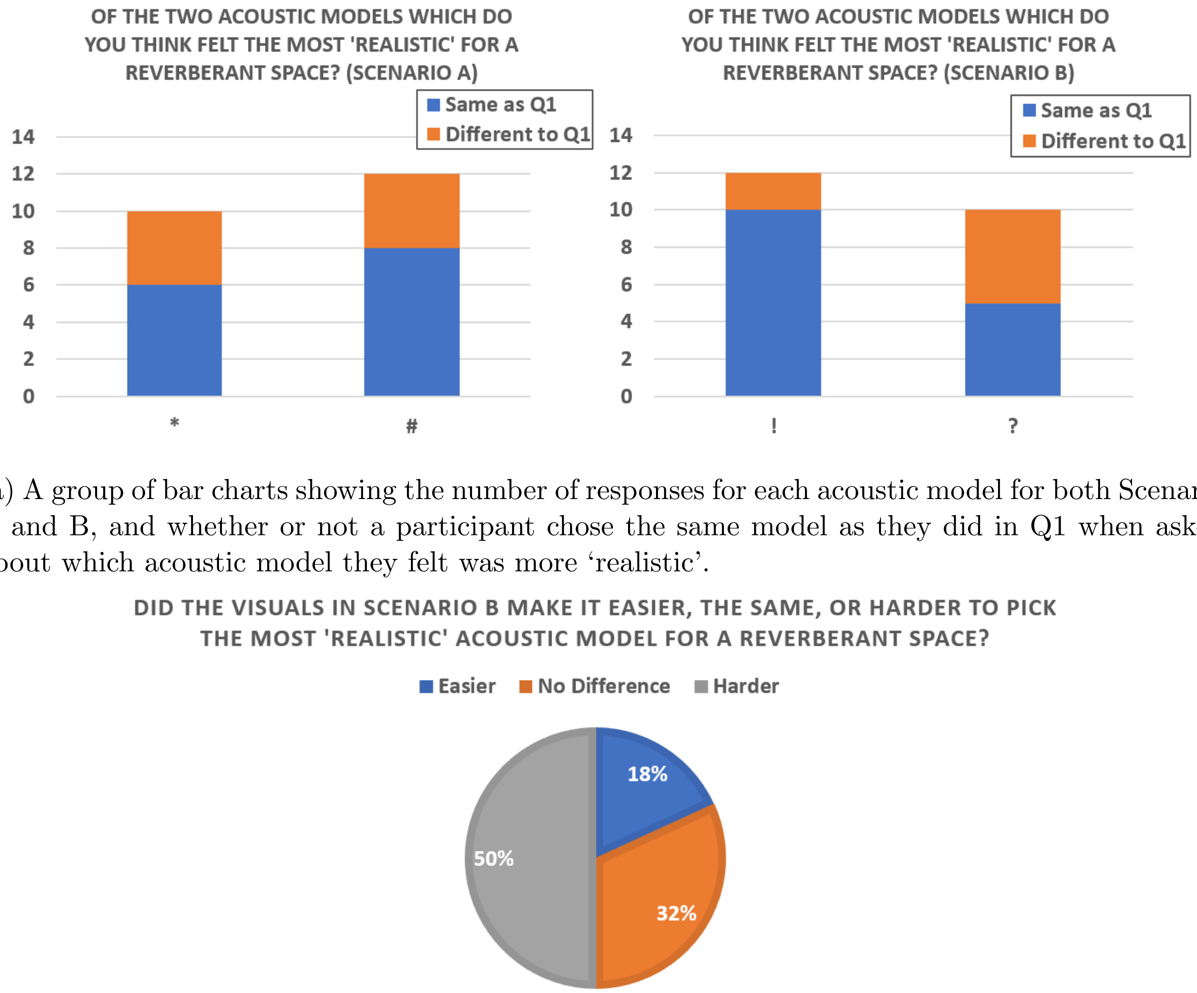
When it came to Scenario B, a higher percentage of responders described their answers as *somewhat unsure* and *unsure*, which led to 36.4% of individuals either giving no reason as to why they chose a model, or talking about how they "couldn't hear any differences" between the two. Again, most of the people who chose the correct model (*?*) described it as "more natural" and full, rather than focussing on negative qualities of the other.

However when compared to Scenario A, a higher percentage of the responses for the generated model *!* (33.3% compared to 25%) based their reasoning on ruling out the other model due to the way it sounded. Comments describing the other model as "too perfect" or "a lot more computer-made" due to how reverberant it was were common once again.

When asked whether the visuals of Scenario B made determining the real acoustic model easier, the same, or harder, the common factor for those who found it harder centred around how disconcerting and off-putting the lack of matching visuals to accompany the audio was. Many participants agreed that looking at a space let them "...make a decision on how you think the space should feel. There was no frame of reference in B". Others went on to explain how the contradictory nature of the scenario led to them feeling dizzy, and that they could not latch onto the sounds as easily.

For the 23% of people that found Scenario B easier, all commented something along the lines of that the simpler visuals helped to make the difference between the two feel "more pronounced" so the differences in the acoustic models were easier to distinguish. There was no correlation between those who found the second scenario easier and their ability to choose the correct model.

Participants were also asked about which model they thought felt the most "realistic" for a reverberant space. Figure 4.11 shows these responses, as well as whether an individual chose the same model as the one chosen for Q1 for each scenario, where they were asked which model they thought was the real-world one.



(a) A group of bar charts showing the number of responses for each acoustic model for both Scenario A and B, and whether or not a participant chose the same model as they did in Q1 when asked about which acoustic model they felt was more ‘realistic’.

(b) A pie chart showing the results regarding the difficulty for determining the answer for Scenario B for which model participants thought was the most realistic.

Figure 4.11: The results for the questions asking about participants’ feelings on which model was the most ‘realistic’ sounding for them.

For this part of the study, participants were not asked about their confidence in answering this question as it did not have a correct answer. It was all about how an individual felt, so their reasons why were deemed more important, hence why the question asking why they made their choice was compulsory (Q5).

For the acoustic models * and #, 8 people swapped to the other model when answering Q4, with a 50/50 split of * to #, and # to *. It is hard to ascertain a pattern as the sample size is small, but all the people that changed to * justified this decision as they felt it had "more echoes" and reminded them more of a church. 75% of those that swapped to the real model rationalised their decision for the same reasons why they had chosen the other model in Q1; # "felt too perfect".

For the models in Scenario B, the majority of participants that swapped model switched from the faked to the real one when asked about which model they thought was more realistic. Even though they didn't think the model was made from real-world data, they thought it felt more real. 50% of the individuals that chose the model labelled ? for Q4 swapped to it, compared to only 16.7% that changed to the generated model. Both of the latter participants justified their change in the same way; "[the model labelled] ? felt more live/natural, but [the model labelled] ! felt more real/realistic". Like with the study detailed in Chapter 3, these results also corroborate research surrounding hyper-realism, where what users believe to be "realistic" and what they believe to be "authentic" are not always the same [34] [171].

When asked if the visuals in Scenario B made it easier, the same, or harder to pick the most 'realistic' acoustic model for a reverberant space, 77.2% of participants chose the same answer as they had for Q1 of the general questions and gave similar justification for their reasons why.

Of the participants who felt differently, the majority felt that the disparity between the visuals and audio did not impact their ability to decide which model felt more realistic, selecting that they felt *no difference* when answering the question. Comparatively, this group all felt that the visual differences did make ascertaining the real model easier or more difficult. For Q1 of the general questions, all made comments pertaining to either how the visual/audio disparity was jarring, or how the lack of a visual placebo made it easier. Whereas for this question, the sentiment was that they felt what they saw had “no impact on what [they were] hearing”. Only 2 participants who changed to a different answer thought the difference in visuals made any kind of impact on choosing which acoustic model was the most realistic.

Ultimately, like with Q1, 50% of people found that the disparity between the audio and visual cues made it harder to gauge differences in the models, with an even smaller percentage compared to Q1 (18% compared to 23%) finding it easier to ascertain.

Only two of the participants were music students, but due to the many extra-curricular activities that take place in the chapel, course studied does not really equate to how much time will be spent listening to music in the chapel, only really indicating deeper familiarity with music historically. When asked what kind of music participants like to listen to, there were a wide variety of answers, with many people just saying “most things” or “anything”, with two responders mentioning classical music specifically. With a sample size this small, it is not likely that there is any correlation to be found between music listened to frequently and whether or not the participant could pick out the real acoustics for the chapel. It would be interesting to run this study again with only people who are very familiar with the acoustics of the space and see how they find the experiment.

4.6 Discussion

These results validate the original hypothesis; sound produced from acoustic modelling cannot be correctly distinguished from sound produced from real-world measurements in reverberant virtual spaces. Thinking on the core aims, it is interesting to see for a second time how many people will hear the true acoustics of a space and decide that they must be false due to the level of reverberation when compared to a less reverberant scenario. What humans perceive as “authentic” is not always the same as what they perceive as “realistic”. Reflecting on this project, the core aim was definitely achieved. A model of a real physical space was recreated in detail and its magnitude captured. The accompanying acoustic models truly demonstrated the fullness of the space aurally.

The VR experience generated has allowed for both current students at the university to explore the chapel in a new way, whilst also allowing people who may have never entered the chapel before to experience it in an aurally authentic way. Due to the nature of the experience itself, participants were limited in their own experiences only minimally. Users could move their whole body both physically within the virtual space as well as virtually through teleportation. Their movements were limited to just the central aisle of the chapel, however that in itself is a large area to explore. The virtual chapel was built at a one-to-one scale, so hypothetically, if one were to set up this experience physically in the chapel, a participant could walk freely in the physical space to explore the virtual one.

4.6.1 Immersion, interactivity and user experience

The immersion imparted by the system was also improved in this study compared to the one detailed in Chapter 3 as, rather than being a video that was recorded at a set height relative to the ground for a participant to watch, it was instead dependent upon the user’s own height.

Participants would have felt as though they were the correct size for their surroundings, matching the player's visual expectation formed in the physical world. However, some aspects were not able to be captured for this specific study. Participants did not have a virtual body, only virtual hands. This was quite disconcerting for some of the participants. However, even with these limitations, participants were immersed in trying to establish which models were the real and most realistic feeling for their environment. This is confirmed through many of the comments surrounding Scenario B, where participants stated they struggled to feel immersed when presented with an environment that did not match what they were hearing.

Prior work recreating real-world spaces for listening to performances has also investigated various methods of interactivity and how they affect user experience. Since this study was conducted, there has been research on interactivity and user experience for other projects revolving around recreating performance spaces aurally. The ruins of Linlithgow Palace were virtually reconstructed at a 1-2-1 scale virtually through the use of laser scanners and a camera, alongside historical evidence to fill in the gaps and furnishings, with the aim of recreating a specific musical performance from history [188]. This would be set up at the real-world location. Initially, the team used a very similar set-up to that of the set-up in this chapter, using the HTC Vive and pre-baked acoustics. In this case however, for practical reasons, the decision was made to move away from the more interactive model in favour of a sit-down, wireless experience. Other people could occlude the infrared sensors, causing the application to glitch; the programme would need to be monitored at all times to run the application for each user; the user could trip on the Vive's wires; and physically moving around a ruin with a VR headset on would most likely lead to the user bumping into other visitors or tripping on the uneven floor, despite the 1-2-1 scale of the model. Therefore, the decision was made to switch to a wireless headset with no physical controller and to make the experience a sit-down one, where the player's avatar movement took place on a 'fly-by-rail' track, taking the user automatically about the space. These changes significantly reduced the agency of the user, however made it safer.

Unfortunately, the Linlithgow Palace experience was not implemented due to the pandemic, so there is no data on whether this sit-down variant of the experience provided the fully immersive effect the team were hoping to achieve. As the study detailed in this chapter was conducted in a fit-for-purpose room that permanently houses the VR equipment used, the issues surrounding the use of the Vive headset were not a problem. No one would be walking around in the room with the user, so there would be no chance of the IR sensors being blocked, or other people physically being in the way; the user had a limited play space to physically move around in, so it was hard for the headset cable to become tangled beneath them; and the researcher was present to set-up the experience for each individual personally. It would be interesting to explore whether having a full body, but being unable to physically walk around would be more or less immersion breaking than having hands and being able to physically walk around, but not having a body. This research is beyond the scope of this thesis.

4.6.2 Audio-visual conflict

At the beginning of this experiment, we had no idea what to expect regarding the impact of the conflicting visual-acoustic experience of Scenario B on a participant's ability to select the real acoustic model. On one hand, there could be a world where the stripped back visuals could enhance the listening experience, allowing participants to hear different nuances in the acoustic models, making the process of deciding which model was real easier. On the other hand, the juxtaposition between what the user was hearing and seeing could instead be disconcerting, reminding the participant that ultimately, the experience was just virtual, and making it harder to ascertain which acoustic model was real. The latter appears to be the case, however, upon looking at the reasoning given by the 5 and 4 participants that found the second scenario easier for choosing the real and most realistic models respectively, some did agree with the initial assessment that the lack of visuals helped them to hear differences more clearly.

4.6.3 Reverberation time disparity

It was also quite surprising to see how little impact an up to one second difference in time for the reverberation times for frequency bands had on this study. There was concern that these differences as seen in Figure 4.8 would be very noticeable, causing people to notice significant differences in the two models. However, this was not the case. Some participants did note that they heard differences between the two models but not "significantly". Some felt one model was fuller, but this was not consistent among participants, with different people choosing opposing models as the full one. On average, when asked to pick out the real acoustic model, 27% of participants could not pick out any differences between them, and ultimately for both scenarios, participants could justify why they thought that the model they chose was the real one. These results once again lean into the discussion surrounding the "reality paradox" discussed earlier [34].

A more accurate representation of the surface materials from the real-world architecture in the virtual chapel could have improved the similarity of the two RT60 frequency bands. An estimation of the materials was made based on what was available in Unity and Resonance Audio, so qualities such as density, or the absorption coefficient could have been different enough from their real-world counterparts to make a significant difference on reverberation time. However, there are other factors in play that could be responsible.

As discussed in Section 4.3.5, some of the claims made by the developers were not repeatable in practice. Resonance Audio comes with two demonstrations to show off the capabilities of the software. In the `ReverbBakingDemo`, users can explore how the reverb probes work in varying acoustic settings, where they already contain the precomputed reverb that has been "baked" into the scene. The documentation states that in these scenes, the probes have been "preloaded with results (RT60s)" [189]. The impression that this gives when compared to the rest of the documentation is that the reverb probes present in the demo were created using the techniques discussed; that the visual materials were mapped with acoustic counterparts and that the probes were placed and baked to result in what a user experiences. However, when these probes were re-baked for a second time, which you would expect to yield the same values, these new RT60s for frequency bands calculated were far less dynamic and less reverberant than the originally provided values. Yet the scene still sounded reverberant enough to feel natural for the provided virtual environment.

These numbers cannot be interacted with by the user in the front end of the software. Therefore, the dynamic, realistic values promised by the software may have been a fabrication to make the virtual environment appear more aurally exciting than the actual result should have been. Potentially, the original demo uses virtual acoustic materials that were not present in the download provided, however that seems unlikely as the other demo included in the same download encourages you to change the material properties of the walls around you, and the same materials subset is provided.

In the demo provided, the difference in the newly baked and pre-loaded reverb probes was significant even though both felt naturally reverberant for the environment. Only the RT60s themselves and the directivity of sound sources placed within the environment are accessible in the interface of Resonance Audio. As the effect of a highly varied RT60 did not have a major impact on the perceived realism of the produced sound, it is likely that the other procedures that Resonance Audio is implementing in the background are much more important for the perception of the sound. For example, the RT60 only directly affects the loss of acoustic power of the sound source over time, but Resonance Audio is also accounting for acoustic phenomena like diffraction, occlusion, and frequency-dependant acoustic losses from reflections, all of which will have a major impact on the resultant sound [3].

Despite the source code for Resonance Audio being available through GitHub, the program itself is not openly clear on the minutiae of how the systems calculating the real-time reverberant properties of a virtual space function. This makes it hard to fully justify its use in future studies of this manner as there are variables that currently are inaccessible to the user.

However, even though the range of the reverberation times for this study were significantly more varied for the live measurements, which made an audible difference between the two models, both were still reverberant. This led to the majority of participants choosing to believe the falsified model was the real one the majority of the time.

4.6.4 Extended exploration

As well as the space for stationary sound sources to listen to in the virtual environment created, an exploration of reactive audio sources, such as the participant having a chance to sing or speak themselves, was something that we potentially wanted to achieve. Investigations into using the built-in microphone in a VR headset to improve immersivity in virtual experiences has been conducted, however, these microphones incur a large lag on direct audio, let alone audio that was filtered through a program to manipulate it. As such, in earlier iterations of potential studies, the ability to incorporate a built-in real-time microphone element had been disregarded.

As mentioned in Chapter 2, there have been a couple of previous studies that have explored the real-time processing of audio inputs and playback, however they involved a separate system to manage that component. Through research for this chapter, it was discovered that Pure Data (Pd), an open-source visual programming environment that can be used to process and generate sound, could be integrated with Unity [190]. Chapter 6 will go into more detail surrounding Pd within this research as a whole. In the case of this study, the use of `LibPdIntegration`, a wrapper for `LibPd` to incorporate Pure Data code into Unity, was explored, specifically its interaction with Unity's Microphone class [191,192]. With some experimentation, a novel Unity microphone class that successfully engaged with the `LibPd` asset and a Pd patch (a unit of code, or program made in Pd) was created, which can be seen in Listing 4.1.

Although it was not possible to get this program working with Resonance Audio at the time, an alternate method of feeding the mic input through a Pd patch that added reverb and delay to my live audio input was achieved. One of the perks of Pure Data is the ability to use it with varying audio drivers. ASIO4ALL is an audio driver which allows for lower latency than regular PC drivers [193]. Pure Data in isolation alongside the ASIO driver allowed for negligible audio lag. Unfortunately, Unity does not allow the use of a different audio driver when running, so as ASIO4ALL could not be integrated, there was still perceptible lag with the mic input. This is ultimately why there was no live audio input feature in this study, however, these explorations are not irrelevant to future work, and helped to inform the research conducted in Chapter 6.

```
using System.Collections;
using System.Collections.Generic;
using UnityEngine;

[RequireComponent(typeof(AudioSource))]
public class Mic_Component : MonoBehaviour
{
  // Boolean flags shows if the microphone is connected
  private bool micConnected = false;

  //The maximum and minimum available recording frequencies
  private int minFreq;
  private int maxFreq;

  //A handle to the attached AudioSource
  private AudioSource goAudioSource;

  void Start()
  {
    //Check if there is at least one microphone connected
    if (Microphone.devices.Length <= 0)
    {
      //Throw a warning message at the console if there isn’t
      Debug.LogWarning(”Microphone not connected!”);
    }
    else //At least one microphone is present
    {
      //Set our flag ‘micConnected’ to true
      micConnected = true;

```

```
            //Get the default microphone recording capabilities
            Microphone.GetDeviceCaps(null, out minFreq, out maxFreq);

            //Documentation says if minFreq and maxFreq are zero, the microphone supports any frequency
            if (minFreq == 0 && maxFreq == 0)
            {
                //so 44100 Hz can be used as the recording sampling rate
                maxFreq = 44100;
            }

            //Get the attached AudioSource component. This is where the ‘lagless’ input is handled
            goAudioSource = this.GetComponent<AudioSource>();
            goAudioSource.clip = Microphone.Start(null, true, 20, maxFreq);
            goAudioSource.loop = true;
            while (!(Microphone.GetPosition(null) > 0)) { }
            goAudioSource.Play();
        }
    }
}
```

Listing 4.1: Microphone script created for Unity that works with Pure Data for lower latency microphone output.

4.6.5 Further research

If this study were to be rerun, there are a few things that would be adapted or changed. Firstly, the internal visuals of the chapel were kept simple, with only block colours being used to represent what is a more visually complicated space. This was done predominately to save time. With further resources, it would have been ideal to be able to overlay the real art and textures from the chapel on top of the model. This model was made to scale; therefore, photos of the space could have been overlayed as a texture within Unity.

LiDAR (Light Detection and Ranging) scanning technologies could also achieve this, potentially just removing the need to have created the model in CAD software to start with. LiDAR scans can be used to quite quickly create accurate 3D models for virtual reality and can be combined with 2D imagery as well to construct photo realistic models of 3D spaces [49–52]. With newer commercial phone and tablet products having inbuilt LiDAR technologies, this method of recreating a physical space virtually will only continue to be more approachable. Users need to just scan a location a single time with this range scanner along with videos and photos of the scene and a 3D representation of this physical location can be created, including both static and moving objects within the scene [53].

Ultimately, there was no access to such a device while creating this study, but it has been shown to be useful in both architectural and environmental mapping studies [194–196] as well generally for VR and AR experiences for fun and health [197, 198] in recent years.

Secondly, although surpassing twenty participants, the sample size was not very large, and it would be interesting to run this study with students who often frequent the chapel or are music students to see if those who are more familiar not only with the acoustics of the space, but with sound in general had a different reaction to the study. It would be especially interesting to explore this with one of the choirs as not only do they rehearse and perform often in the chapel, they also do so in a group, like with the performance in the study.

Finally, initially it was hoped that in this study, there would be the ability to have live audio inputs interacting with the acoustic models, not just pre-recorded ones, but for reasons stated in section 4.6.4, this was not integrable within the timeframe and would have required a separate system to handle it. With the assumptions made in the study approximating the impulse response for the main corridor of the chapel as equal throughout, these two IRs could have been used with independent software such as Pure Data and have had live audio fed through. However, interchanging between the two in time with the VR buttons would not have been trivial, especially as it would be run on a separate device. The study would have to be reworked and may not have been as dynamically responsive, resulting in a less smooth real—time experience if all audio was to be controlled together, and not necessarily from within the VR scenes. However, as will be explored in Chapter 6, utilising off-ear headphones with surround speakers, two separate systems could be run dependently alongside each other.

4.7 Summary and conclusions

This chapter outlined the process of creating a real-world space virtually for VR with accompanying acoustical properties that were either a computer modelled approximation or a real-world measured model of that space. Utilising CAD software alongside Resonance Audio and Unity, a model of Royal Holloway's Chapel was created for VR and could be explored visually, and aurally for a 4-part choir and female speaking voice.

Participants were tasked with deciding which of the two models was real, and which felt most realistic for the reverberant space. It has been shown that sound produced from acoustic modelling cannot be correctly distinguished from sound produced from real-world measurements in reverberant virtual spaces, despite the two sounding different. Indeed if anything, people are slightly more inclined to believe that, when directly compared, a less reverberant, generated acoustic model is the most realistic of the two.

Section 4.5.3 also demonstrated that when presented with a conflicting visual-to-audio environment, 50% of participants found it harder to discern which model was real, when compared to the scenario where the visual and audio environments matched. These results were explored in Section 4.6, alongside a discussion of Resonance Audio and its suitability for future experiments aurally exploring real-world spaces virtually. A novel microphone class was created to be used with the `LibPdIntegration` wrapper for Pure Data to incorporate real-time, user audio stimuli within the virtual acoustic environment, however without a suitable lower latency audio driver, this class was not suitable for this study due to too high a latency. Finally, future improvements for the experiment were discussed, exploring the concepts of using photography or LiDAR to make the 3D model of the Chapel more visually accurate.

Chapter 5

The Design and Implementation of a Real-Time Audio Convolution System

> *"We seek the same feeling from a psychologically immersive experience that we do from a plunge in the ocean or swimming pool: the sensation of being surrounded by a completely other reality, as different as water is from air, that takes over all of our attention, our whole perceptual apparatus."*
>
> - J. H. Murray, *Hamlet on the Holodeck: The Future of Narrative Cyberspace* [25]

There has been a running theme throughout this research thus far: people comparing two sounds where one is audibly more reverberant than the other and deciding whether that makes the sound more, or less, realistic to them. To people who don't study music or acoustics, the elements that define the "echoey-ness" of a sound may not be obvious. When defining something as more echoey, more reverberant, what do we actually mean? Is it purely the reverberation time alone, or do other factors play a part in influencing a person's opinion of the acoustics of a reverberant space?

For someone who played music from a young age in vibrant acoustic environments, these questions may be hard to personally answer. Engaging with people who have never even set foot in a church-like building before and have no concept of what any audio is like in these spaces, has been interesting to explore. So further researching which aspects of acoustics people latch onto when imagining what a space should sound like, if they provide it with an impulse, would allow us to understand how to better authentically recreate those spaces virtually.

These final chapters are where these questions shall be examined. How important are the specific acoustic properties of a space when it comes to impersonating real-world acoustics? What is the range for which factors such as the length of the reverberation tail, or the Initial Time Delay Gap (ITDG), can be extended or diminished and still feel like the true results? This is an area of research that is intriguing to explore, especially when it comes to the potential differences between what the average person can pick out compared to what people who regularly perform in reverberant acoustic environments distinguish as the key properties.

This chapter describes the design and implementation of a novel, real-time audio convolution system which will allow for the real-time manipulation of a live input from a performer, granting them the ability to alter the virtual performance space in which their voice is being simulated in. The system created allows the user to change acoustic variables of an impulse response as it is convolved with their own voice through a portable, light-weight device named “The Bela Box”.

The chapter begins with a discussion of what The Bela Box needed to achieve for it to function in the intended manner. Section 5.2 details the use of Pure Data for creating a novel real-time convolution patch for varying impulse responses and audio inputs, followed by a section on the use of Bela as a tool for managing this low latency audio manipulation and why ultimately Pure Data became inappropriate for this task. Section 5.4 explores audio processing, looking more specifically at how real-time audio programming can be implemented as well as the limitations of these systems. This section also details how through a mix of frequency domain and direct form convolutions, C++ and Bela can be combined to create a low-latency convolver. Section 5.5 looks at future developments for The Bela Box, followed by a summary of the chapter.

5.1 Concept for the real-time convolution system

5.1.1 Desired outcome

The aim is to create a device that would be of use in future research investigating how true to reality an audio simulation of a real environment needs to be for a person to feel as though it is true to the physical location when they provide a live input via singing. It also provides a method for performers to get a feel for singing in a specific real environment virtually through just an impulse response of that location.

5.1.2 Real-time convolution

As described in detail in Section 2.6, convolution of a room impulse response with a sound source can create the effect of that sound coming from the environment of the RIR. There is nothing stopping this from being a real-time effect, as long as a low enough latency system is used to manage it. This would allow for exploration where the sound source could be somebody singing or speaking live. This also allows for the RIR of a real space to be measured and implemented into the system, so a performer hears themselves in the virtual version of a real physical space.

5.1.3 Real-time manipulation

With a low enough latency, a performer can not only experience the simulation of themselves performing in a virtual version of a real environment, but alter the parameters of the acoustic variables of an impulse response. Increasing or decreasing the ITDG, or the reverberant tail while singing. Or fully changing the RIR being convolved to a different one.

5.1.4 Usability

It is intended for this device to be used in tandem with a visual virtual environment through a VR headset, so the device needs to be usable without necessarily being visible at the time of use.

5.2 Real-time singer voice feedback

There are multiple programmes that can integrate real-time convolutions with live audio inputs, including Pure Data. Pure Data is a visual programming language for real-time audio, video, and graphical processing. These programmes, or patches, can be controlled by other patches, allowing for programmers to create an interface through which their code can be altered within its intended bounds.

5.2.1 Pure Data patch for real-time audio manipulation

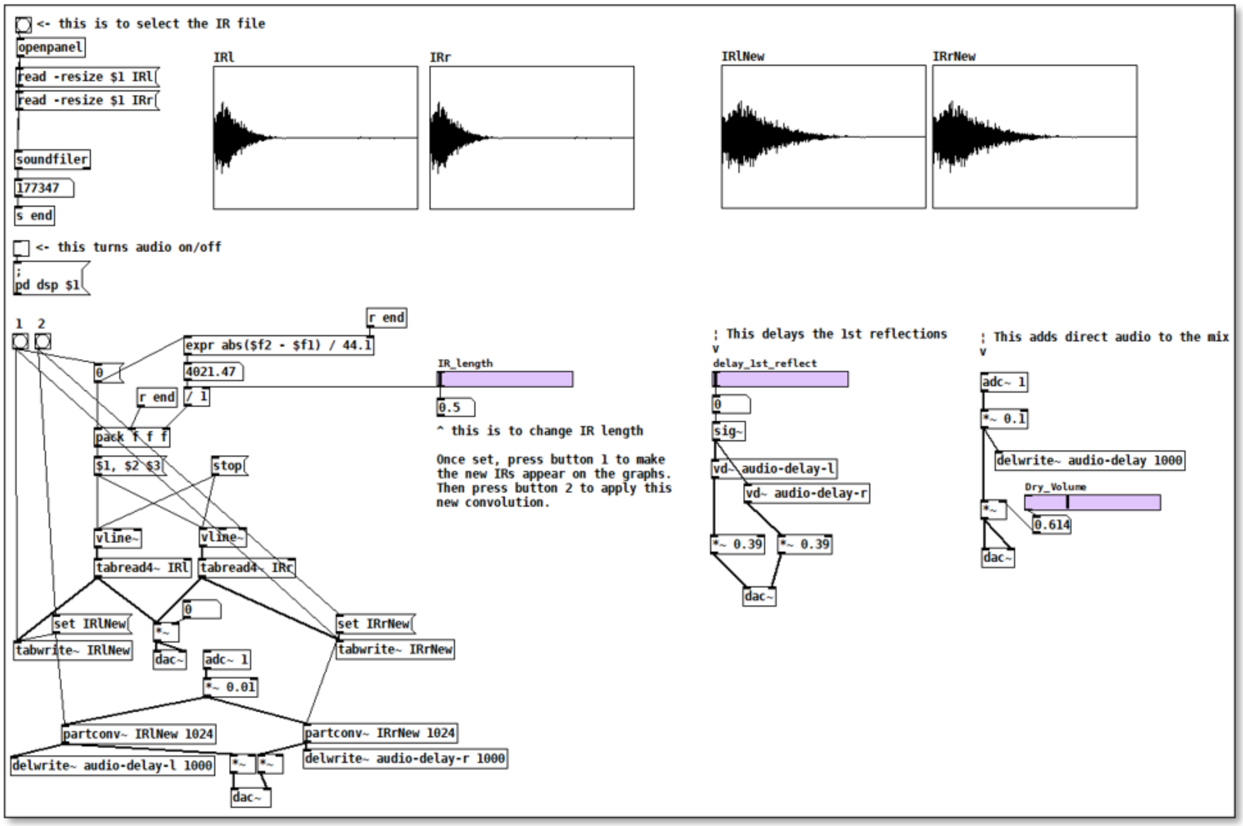
A Pd patch was created that could, in real-time, alter the way a person's voice sounds via the manipulation of three sliders on an interactive panel. As this was a prototype, the process and its interface were all contained in one patch. Had this programme been intended for use by other people, the interactable sliders of this patch would have been separated onto a second patch to indicate more clearly how to operate the software and hide the development end of the code.

The patch allows the participant to play with the lengths and volumes of various sound cues such as ITDG. Figure 5.1 shows the original patch that was created for this project.

The patch is labelled at the parts where a user would interact. To start with, an IR file can be selected from the device and read into the patch. Using "send" and "receive" objects (`s` and `r`), this IR array can be sent elsewhere in the patch. This original IR array is split for headphone usage into a left and right signal labelled `IRl` and `IRr` respectively. These impulses can be seen in the leftmost pair of graphs in Figure 5.1. For the auralisation, the direct sound should not be included in the impulse response file, as that comes directly (in-head) from the speaker/singer. Therefore it is necessary to edit the impulse response to remove this direct sound component.

`expr` is used to find the length of the IRs, allowing for the first of the sound cue manipulations. The `IR_length` slider changes the length of the IR and reverberant tail, ranging between halving and doubling the original impulse response. This change is then read by the patch, and this new IR is saved and set as `IRlNew` and `IRrNew`.

This patch works primarily through the Pd external object `partconv~`. `partconv~` implements partitioned fast convolution that can take in long impulse responses for reverb, convolving input signals [199]. The external takes in these new IR arrays and partition size and outputs a convolved signal. The help patch demonstrates it being used for both live



Once set, press button 1 to make the new IRs appear on the graphs. Then press button 2 to apply this new convolution.

Figure 5.1: A novel Pure Data patch that, in real-time, can alter the way a person's voice sounds through the manipulation of 3 sliders; IR length, Initial Time Delay Gap, and direct audio volume.

audio input as well as other signals, such as a sine wave and Dirac impulse. We take in a live signal through `adc∼1`, and it is outputted through `dac∼` once the convolution has been applied.

The user also has the ability to shift the delay of this signal using the `delay_1st_reflect` slider. This is achieved through the use of `delread∼` (or equivalently, `vd∼` which is used in this patch) and `delwrite∼`. The first argument of each of these objects is for the name of the delay line created by `delwrite~`, and the second argument is the length of the delay line in milliseconds. The delay time, applied to `vd~` using the slider, cannot exceed this value, which for this patch is 1000 ms. This changes the length of the ITDG.

This patch is intended to be used with off-ear headphones or speakers so that you are able to hear your own voice naturally. However, for those who don't have access to this kind of set-up, direct audio can be added to the mix through the right-most slider labelled `dry volume`.

Finally, there are two buttons built into the patch that are labelled `1` and `2`. To input the change in IR length, the user must first press button one, wait a short amount of time for the patch to update, which is signalled by the two `IRNew` graphs, then press button 2 to apply the new convolution to their voice.

The design of this patch allows for a user to change the values of the three sliders to what they think the real values for a room impulse response of a given space is. Pure Data allows for patches to refer to other patches, so if Figure 5.1 had been the final iteration, the actual values would have been obfuscated, with the user seeing an interface with just sliders that pull the information from the original patch. For the length of the reverberant tail, the value should be 1, as it is 1 multiplied by the original IR. Calculating the correct first delay time is a little trickier.

5.2.2 Measuring the latency of a system

The key to measuring the latency of a system is to understand the ITDG (initial time delay gap) between the direct sound and the first arriving reflection. To do this, we need to ascertain specific values first:

1. the time between the direct sound and the first reflection
2. the system latency of Pure Data
3. the delay created from `partconv~`
4. the latency of the recording and listening setup

Time between the direct sound and the 1st reflection:

Through analysing the waveform in Audacity, the time difference between the first reflection and the original impulse was measured to be 32.17 ms.

System latency of Pure Data:

This is set in the *Audio Settings* in the *Media* tab. Using ASIO4ALL, a low latency audio driver, as the audio driver, this value was set at 8 ms, which was the lowest setting possible before the audio inputs became distorted.

Delay from partconv~:

The help patch for `partconv~` details the latency you can expect from using it,

"The partition size must be a power of 2 greater than or equal to the block size. Larger partition sizes are more efficient, to a point, but increase latency (the delay between input and output is equal to the partition size minus the block size)." [200]

The partition size and block sizes used in this patch were 1024 and 64 respectively, giving us a difference of 960. From this, we can calculate the latency as,

$$\text{Partition Size } - \text{ Block Size} = \text{ Delay}$$
$$1024 - 64 = 960 \ \textit{samples}$$
$$\frac{\text{Delay}}{\text{Frequency}} = \text{ Latency}$$
$$\frac{960}{44.1kHz} = 0.021769s = 21.77ms.$$

Latency of recording setup:

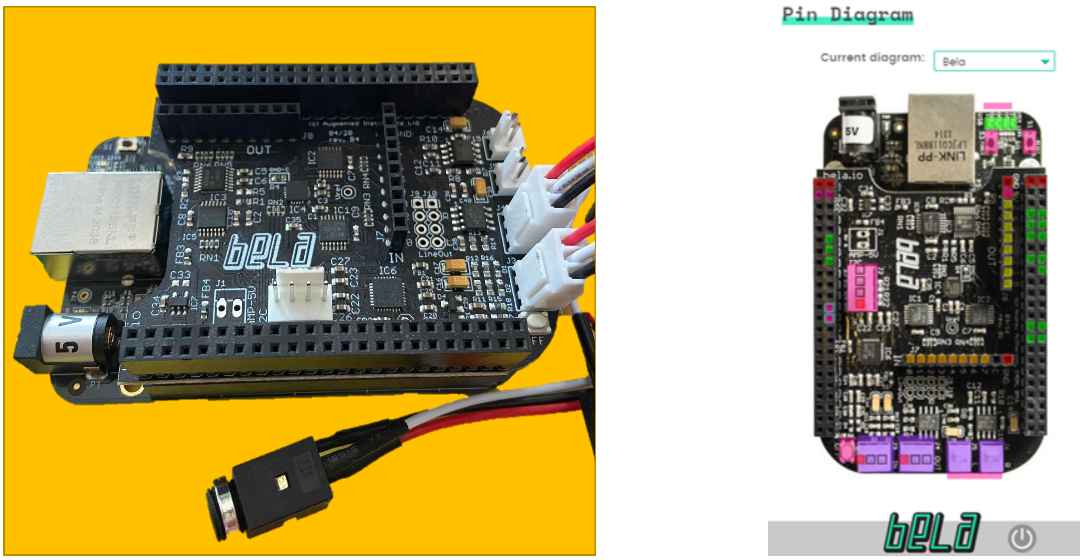
A microphone was placed at the tweeter of the headphones, and a test sound was simultaneously played back and recorded in REAPER. The temporal offset of the recorded sound was compared to the original, giving a latency of 0.37 ms.

Therefore, the total latency incurred on the system by the whole setup is 30.14 ms. This can be reduced by changing the partition size in `partconv~`'s first argument. However, this was not required as the measured time between the direct sound and 1st reflection was 32.17 ms.

5.3 Bela and ultra-low latency

As mentioned in Chapter 4, to use the ASIO drivers alongside VR software, we would need to use two separate systems/computers as Unity and Unreal Engine do not work with ASIO. This would also make the interfacing with the patch by the participant very clunky during an experiment, as they would be required to remove their VR headset, then go to another screen to vary the slider values, then re-enter VR.

Bela is an embedded computing platform aimed at interactive projects due to its ultra-low latency and high-quality audio packed into a small, self-contained device. It is designed



(a)

(b)

Figure 5.2: (a) A photo of the standard Bela Board with audio adapter cables attached for use with headphones and a microphone; (b) A pin diagram of the Bela board taken from the Bela local IDE [10].

with audio in mind, running a custom Linux audio environment that can give buffer sizes as small as two samples, meaning that for audio in, to audio out, a latency as low as 1 ms can be achieved [201]. It also boasts 8 analogue inputs and outputs, 16 digital I/O pins, and can save code to run without being connected to a computer. Figure 5.2 shows the Bela board used for this project alongside a virtual pin diagram for the device.

5.3.1 Bela and Pure Data

The device was chosen for these above qualities, along with the fact that it is compatible with Pure Data, meaning that an independent device could handle the convolution of RIRs with a live audio input, as required. The patch required some alterations to make it fully functional with the Bela board.

The first thing to note is that any externals used with Pure Data need to be uploaded onto the IDE. Bela uses the *libpd* library, an embeddable library for Pd. The Bela IDE is not a Pure Data patching environment, you cannot edit the patch there, instead you upload them saved as `_main.pd`. This is the first major difference when compared to creating patches on a computer.

Next, the sliders on the patch were reworked as potentiometers using the analogue pins on the device. These changes were a simple one-for-one rework for two of the sliders, directly swapping the digital process for an analogue one. However, due to different variables in Pd accepting specific types of message or connection, the impulse response length slider was harder to implement as the old method would not translate directly over. Linear potentiometers were purchased for this project as it was more appropriate to have a linear range for each variable compared to an exponential one, allowing participants to fully explore the range of audio possibilities.

The original patch required buttons to engage with the new impulse responses created, therefore simple analogue versions of these buttons were acquired to be implemented. Finally, to use only the on-board 3.5 mm microphone jack, a microphone with its own power source is required for integration with the Bela board. The board can power a microphone if necessary, but this is a more involved process that cannot simply utilise the provided jack, instead requiring soldering to the board itself. Unfortunately, using Pure Data for this project was ultimately untenable. Figure 5.3 displays the code in its final, incomplete form.

Whenever this patch was run, the board would crash, and the audio input through the microphone returned crackly and disfigured. Underneath the crackle, there was also a repeating clicking at the top end of the range of the potentiometers. After much research, a solution could not be obtained.

Through this, it was highlighted by other Bela users that Pd was not the most suitable

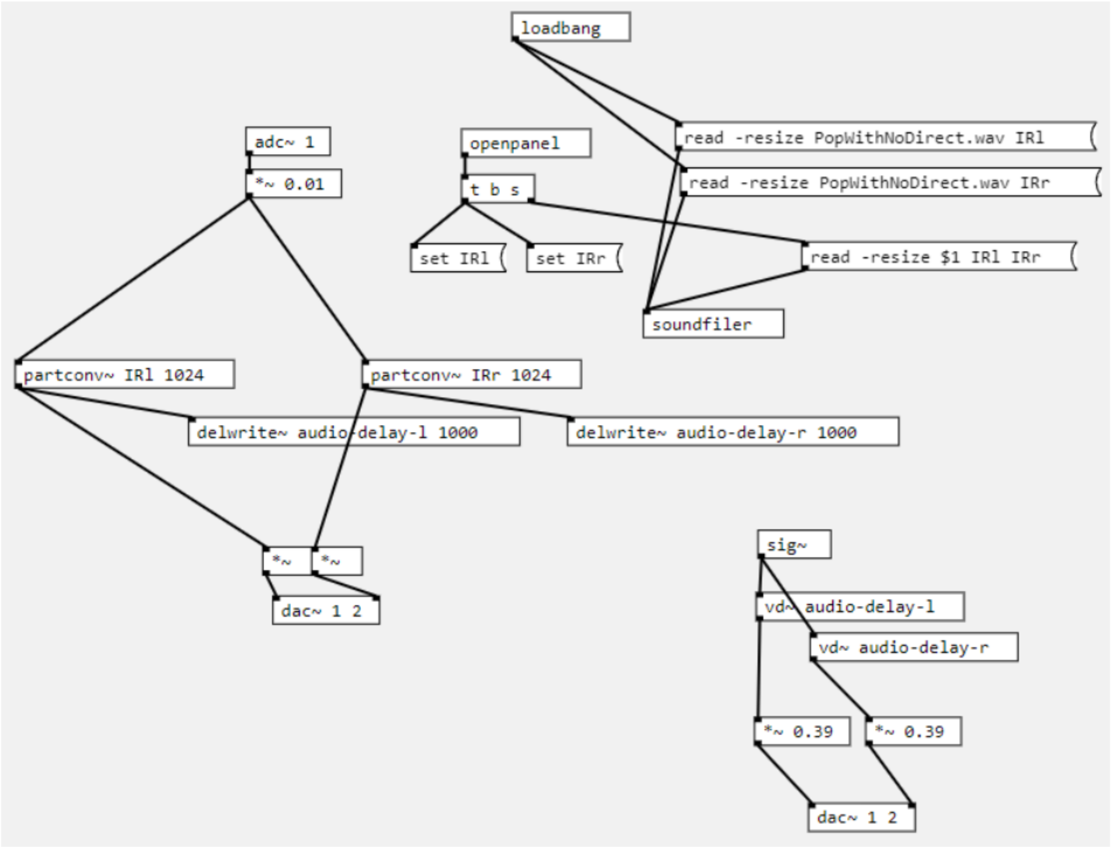


Figure 5.3: The Pure Data patch reworked for the Bela board that was ultimately discarded.

program for this specific project, and that through C++, a solution to the problems would likely be found.

5.4 Real-time convolution with Bela in C++

In Chapter 2, direct form and frequency domain convolution were discussed, and their limitations when used in real-time audio were highlighted. Bela comes with a pre-built real-time convolution class and documentation, with instructions on how to engage with a high-performance mode for larger IRs. However, this convolution was not appropriate for this project. The Bela class performs a time-domain convolution of a signal with an impulse response, meaning that for larger values of N, the system struggles to run. This convolver render was capped at an IR length of 12000 frames, which was not long enough for the chapel, and caused the IDE to become unresponsive.

Fortunately, research has been done into how one could combine frequency domain and direct form convolutions to create a lower latency hybrid convolution [103] [104] [202]. This is especially useful for this case where the IR file is long.

5.4.1 Convolver creation

The design of this convolution implementation is built from the work of Steinmetz, which is heavily inspired by the technique proposed by Gardner, referred to as the 'zero-latency' convolution [101, 202]. Signal processing will always incur some small latency due to hardware, however the applied approach only imparts latency through executing the computation, unlike in block-based convolution where further latency is imparted while samples accumulate in a buffer [104]. This approach combines both the direct form and frequency domain convolutions in an attempt to reduce latency and computation complexity. For this configuration, the impulse response h is split into multiple blocks (as in the block-based frequency domain convolution) of increasing size, as seen in Figure 5.4.

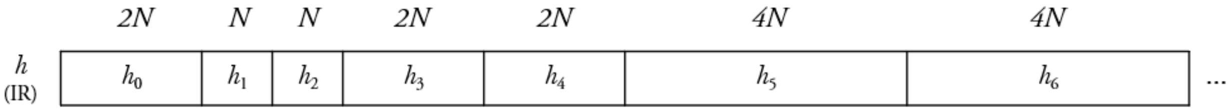


Figure 5.4: A visual representation of how the impulse responses are split into blocks of varying sizes, where h is the impulse response and N is the block size for which the FFT convolution becomes more efficient than the direct form convolution.

The block sizes differ to achieve the best balance between latency and computation complexity. The first block is implemented with a direct form convolution to capitalise on only requiring a single input sample to process the next output. It is set to a block size of $2N$ samples, where N is the block size for which an FFT convolution becomes more efficient than the direct form. After this first block is processed, separate tasks are scheduled to compute the latter blocks after enough input samples have been collected. These tasks are ordered in decreasing priority, with the smaller earlier blocks being processed first as they are required sooner than the larger later ones.

Three optimisations stated by Gardner for convolution are implemented in this approach; precomputing the spectra for all blocks of the filter h, utilising a real valued FFT to exploit symmetry, and exploiting symmetry in complex multiplication [202]. This system utilises a combination of direct form and frequency domain convolution implementations to reduce latency through the use of two lower-level classes, `DirectConvolver` and `FFTConvolver` respectively [101].

DirectConvolver

The `DirectConvolver` class is the simplest of the two, utilising the direct form convolution. As mentioned above, this type of convolution is quite efficient for a small number of coefficients, however it becomes more and more costly as this number increases. Therefore, the direct convolver is just going to be used to process the first block of the filter. This operation is set in motion during the initialisation of the class where only this first block is considered and its filter coefficients are considered within the convolver class.

FFTConvolver

Although more complex, this class shares many of the same key programming elements as the direct convolver class, including requiring a vector containing filter coefficients during initialisation. This class however exists to convolve all but the first impulse response block. Also, the supplied FFT size must be twice as big as the number of samples in the supplied filter. In `setup()`, Bela `fft` objects are created and set up; `fftX` for the current input block and `fftH` for its associated filter block. An FFT buffer is also instantiated for storing the results of the complex multiplication using these blocks (see Section 2.6.2 for the mathematical expressions).

ZLConvolver

The `ZLConvolver` class combines the above lower level classes to create a complete "zero-latency" convolution system. The most crucial stage of this procedure is the division of the full impulse response file into a set number of blocks which each have their own convolver. Following the pattern set in Figure 5.4 where FFT size is always twice the block size, enough samples are then read and a new convolver instance is created for that block. A single `DirectConvolver` instance is initiated for the first block h_0, whereas for subsequent blocks, `FFTConvolver` objects are stored in a vector for them.

5.5 Audio processing programme

By combining this adapted, low-latency convolver with existing Bela classes, an audio processing programme was designed to recreate the original concept explored by the Pure Data patch from Figure 5.1, but more efficiently and in a way through which the benefits of Bela could be utilised.

There are two key facets to the program; the part that controls the manipulation of the convolution, and the part that controls the delay. Note that as a starting point for the design of these parts of the code, the lecture series co-developed by Bela and Queen Mary University of London were referenced. Specifically Lecture 20 and Lecture 11 on phase vocoders and circular buffers respectively [203, 204].

5.5.1 Key differences from the Pure Data patch

The original aim, as achieved through the computer-ran Pd patch, was for the user to upload a single impulse response whose ITDG and length could be manipulated continuously and freely in real-time while intaking a live audio input. Unlike with the original Pd patch, it was not possible to create a slider that could vary the length of the impulse response in real-time as it caused the board to crash as the CPU usage was too high. There may be a method of creating this more efficiently so that it would function without crashing, however this was not discovered during research for this project.

Instead, five IRs with varying lengths were created and saved in advance using the original Pd patch and original IR seen in Figure 5.1, but with an added save function. The original IR was pre-manipulated using the original Pure Data patch to create 5 variations of it, one of which was the real impulse response to be uploaded onto the Bela board. This replaces the continuous manipulation of length that was originally achievable through the computer Pd patch. These were then loaded onto the board and could be called by the convolver classes. Five files were chosen as that was the maximum amount that could be stored on the board before it crashed on boot.

Before the setup function, `ZLConvolver` is called and the number of convolvers equal to the number of impulse responses that are being applied are added. The IR files themselves are then loaded. `gProcessInput` is then set to true, allowing live input to be processed, which is followed by a line assigning which input channel to process. A circular buffer and read and write pointers are initialised for implementing delay.

5.5.2 Real-time audio processing

Real-time audio is processed in blocks. As an audio signal arrives at an input, it is gathered up into an array, or buffer. This buffer is then passed into the render function, which will apply whatever effects are being applied to the incoming audio, while the next buffer is filled. Once the audio computation is completed, this initial buffer is sent to the hardware to be played back while the next buffer is processed. This means that at any given time, there are three processes happening: the hardware is gathering the audio samples into a buffer; the render function is processing the buffer, applying some effect; and finally, the audio output hardware is playing back a buffer. This structure highlights the latency of a system being equal to twice the buffer length plus any extra latency inherent to the written processing code [205]. Figure 5.5 demonstrates this visually.

If the intention is to store these buffers to apply some effect to the samples within them, for example, to delay them by a set amount of time, the stored samples within the buffer need to be updated every time the render function is looped through. This is where circular buffers are useful. In the setup function, the circular buffer is allocated and the convolvers are configured.

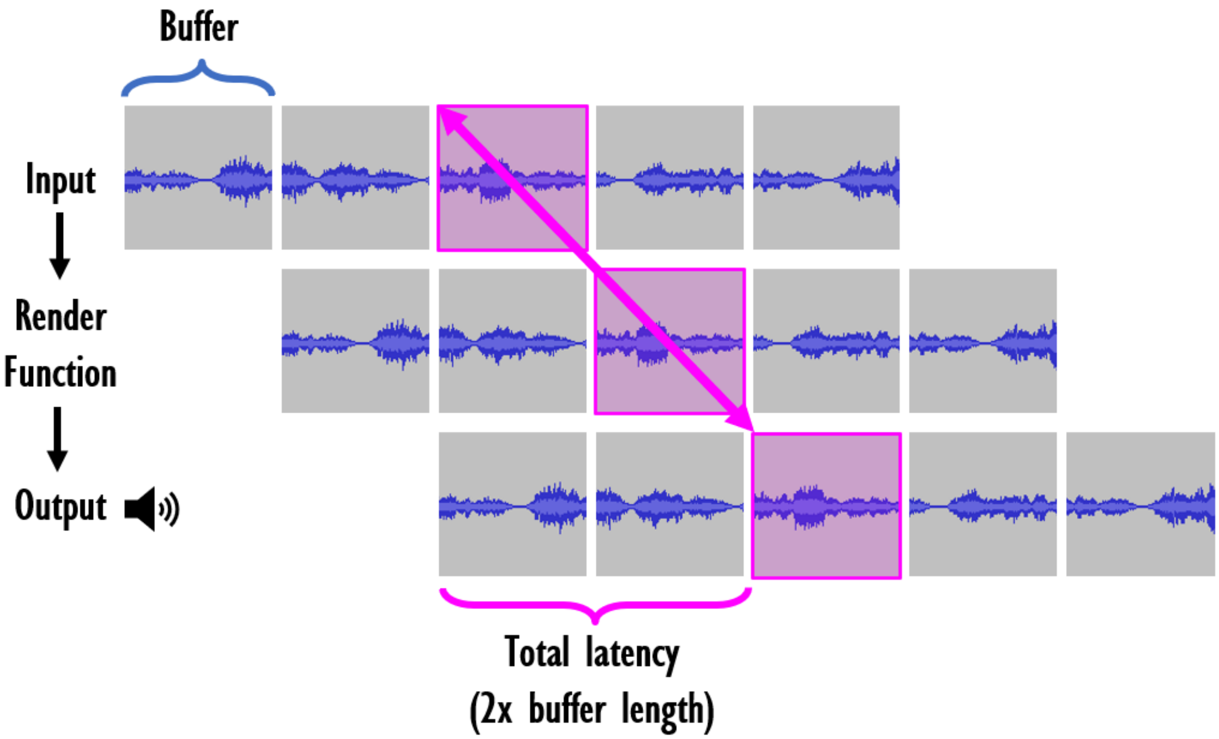


Figure 5.5: A visual representation showing how latency is incurred in a system. Each block of audio is a buffer, and the pink highlighted buffers show why the latency of a system is at minimum twice the buffer length.

5.5.3 Circular buffers and pointers

Circular buffers are a type of memory buffer (array) that acts like a loop when updating stored samples within a buffer. The buffer ultimately needs to store the N most recent samples. However not all memory buffers do this efficiently. Picture this buffer as a movable belt that can save the last 8 samples for a 50 sample system, for example. When the 9th sample needs to be stored, some memory buffers will just add it to the front of the belt, moving all the previously stored samples to replace the oldest one. See Figure 5.6 for a visual explanation. This is inefficient as the stored samples are all being moved every step, which is wasteful. Instead, a circular buffer can be implemented using read and write pointers.

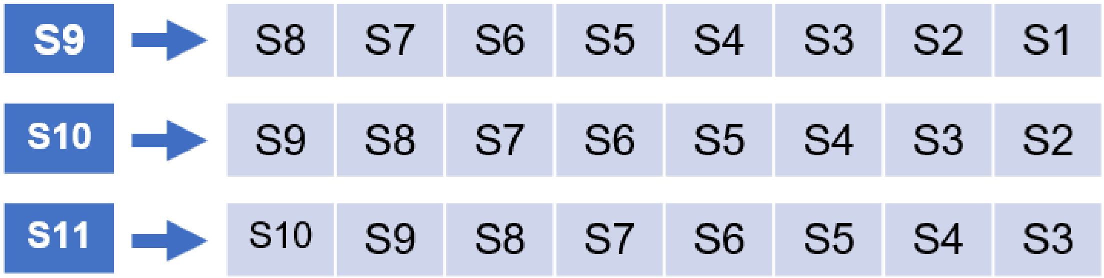


Figure 5.6: A visual representation of an inefficient buffer system where every stored sample is moved every time a new sample is added.

This is done by leaving the old samples in place when a new one is added and instead just directly replacing, or rewriting, the oldest sample. The write pointer dictates where a new sample is written in a buffer and once it reaches the “end”, i.e. once the buffer is completely full, the write pointer can be wrapped around back to the beginning to write over the oldest samples. This can be visualised by imagining this process as a doughnut, which can be seen in Figure 5.7.

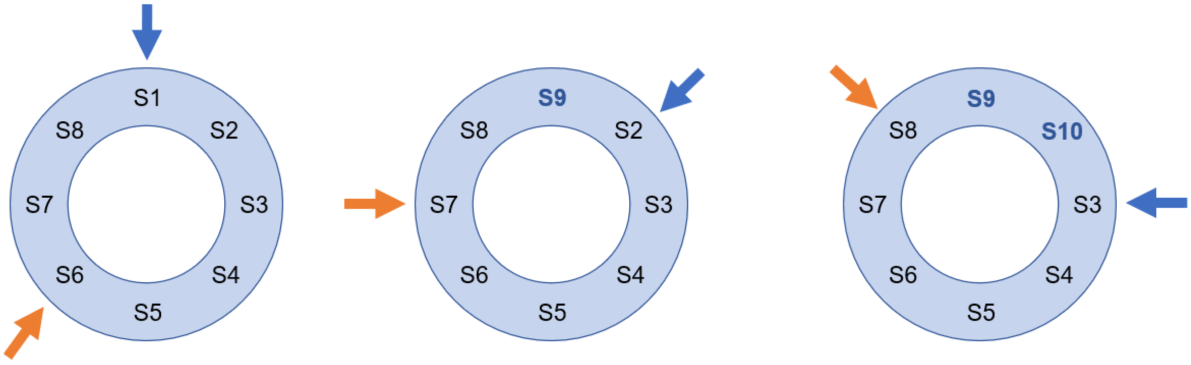


Figure 5.7: A visual representation of how circular buffers work. The blue arrow represents the write pointer, and the orange arrow the read pointer.

Earlier samples can be found by looking backwards (or going anti clockwise) from the write pointer. This can be achieved through some simple arithmetic calculated every audio frame (sample) to read samples out at a particular delay. Or a second pointer called a read pointer, could be implemented in the same fashion as a write pointer to keep track of what is being read. This method means that the amount the outputted signal can be delayed by is dictated by the distance between the two pointers as opposed to the buffer size. By moving the read pointer, the delay incurred on the outputted audio can be manipulated in real time. It is in the setup function that the circular buffer is allocated and the convolvers are configured.

5.5.4 Render function

Convolver process

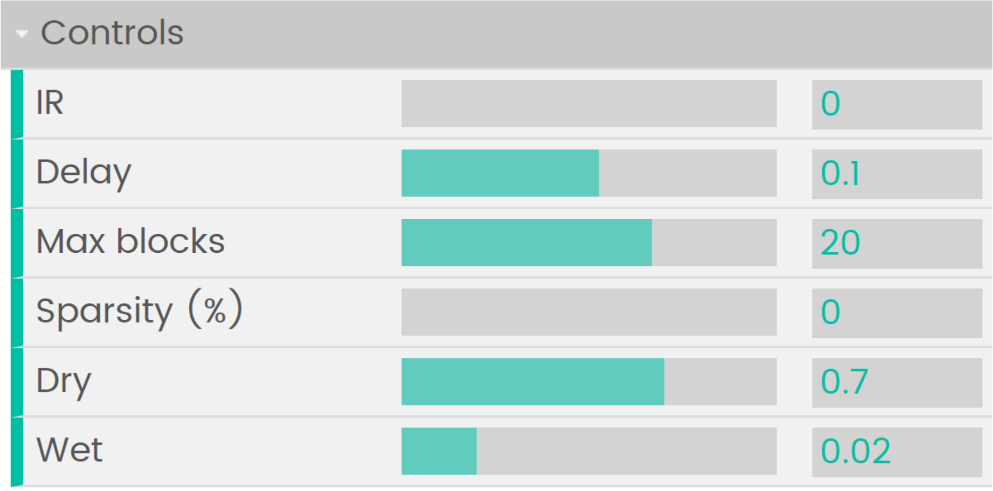
In the render function, a switch statement is implemented to allow the different convolutions to be cycled through by the user. These convolver functions take in the incoming signal from the microphone to be processed and are assigned to be the new outgoing signals.

Delay process

This outgoing signal, named `out`, is the signal that needs to be delayed. At the beginning of the render function, the read pointer is calculated based on the location of the write pointer by taking the write pointer and subtracting from it the desired delay of the output in samples, adding a multiple of the buffer size, then taking this as a percentage of the buffer size. Then, once `out` has been convolved, the buffer is overwritten at the write pointer and the output, named `outDelay`, is read, which is the delayed signal. Finally, both the pointers are incremented and wrapped, as is required for a circular buffer.

5.5.5 Graphical User Interface

To ensure the program was working before implementing a user interface with physical sliders, a GUI was created, as shown in Figure 5.8. This virtual interface allows a user to manipulate the convolution algorithm and other facets in real time.



Controls

IR 0

Delay 0.1

Max blocks 20

Sparsity (%) 0

Dry 0.7

Wet 0.02

Figure 5.8: The Graphical User Interface (GUI) for a Bela patch that enables user interaction with the convolution algorithm in real-time.

The first slider, labelled *IR* lets the user switch between the multiple, pre-loaded IR lengths. These IRs are all scaled versions of the original impulse response, but set as discrete values instead of continuous ones like the Pure Data patch, which allowed for an individual to apply one of thousands of impulse responses created by a slider. This program loads set IR files saved to the Bela board due to the inability to create a functioning copy of this Pd slider in C++ that maintained a low latency performance while running that would not crash the board.

The slider labelled *Delay* controls the delay of the outputted convolved signal. It alters the length of the Initial Time Delay Gap, shifting the first significant reflection along with everything that comes after it. This slider ranges from 0.00 ms to a 0.20 ms delay on hearing the first reflection, which can be freely manipulated while the program is running.

Two sliders labelled *Max blocks* and *Sparsity* are implemented as part of the `ZLConvolver` class and exist to enable control over the number of blocks of the convolution filter applied. *Max blocks* does this by excluding blocks from the end of the IR, whereas *Sparsity* progressively removes blocks throughout the entire filter. These exist to aid with computational load if required when starting up the program.

Finally, there are *Dry* and *Wet* mix sliders which control the volume of the direct microphone input and convolved signals respectively.

The Bela board allows for remote running without a PC by changing the settings of the device to load a project on boot, only requiring an external power source, such as a portable battery pack. By having a physical interface not on a screen, the program could be utilised in settings where a user could not see the screen, for example, within a virtual reality experience where removing the headset could interrupt immersion.

5.5.6 Physical interface conversion

The base code outlined in the beginning of this section is still implemented for this version of the program, however the custom GUI function is partially replaced with code supporting analogue inputs through 10 K potentiometers. The user is provided with a set of dials to control the *Delay*, *Dry*, and *IR* sliders physically.

Similarly to how slider inputs were mapped for the GUI in the `setup()` function, the ADC inputs from the potentiometers are declared and mapped to a useful value range based on their usage in the `render()` function. The input range for the potentiometers is from 0 to 3.3 / 4.096, corresponding to a reading range of 0 to roughly 0.806. This is because an input voltage of 3.3 V is being used. Part of this code can be seen in Listing 5.1. The full project can be found on GitHub [206].

```cpp
void render(BelaContext *context, void *userData)
{
  ...
  for (unsigned int n = 0; n < context->audioFrames; n++)
  {
    // Declare ADC inputs from potentiometers
    float input0 = analogRead(context, n/2, 0);
    float input1 = analogRead(context, n/2, 1);
    input1 = input1Filter.process(input1);
    float input2 = analogRead(context, n/2, 2);
    input2 = input2Filter.process(input2);

    // Map ADC inputs to useful value range based on usage
    float amplitudeDB = map(input0, 0, 3.4 / 4.096, -40, -6);
    float delay = map(input1, 0, 3.4 / 4.096, 0, 0.40);
    float room = map(input2, 0, 3.4 / 4.096, 0, 2.9);

    // Further value adjustment
    float dry = powf(10.0, amplitudeDB / 20); //convert dB to linear amplitude
    ...
  }
}
```

Listing 5.1: Part of the Bela code which declares, maps, and filters ADC inputs

This variation of the board and its code were installed into a device which shall be called The Bela Box, with three potentiometers, easy access to the headphone and microphone jacks, and an easily accessible power cable. The potentiometers allow for participants to adjust the length of the IR file being convolved, the volume of direct audio, and ITDG. The dials are all next to each other and easy to feel as the intention, as will be seen in Chapter 6, is that participants will be altering these values while wearing a VR headset, and therefore will not be able to see the box. A circuit diagram and a photo of the Bela Box can be seen in Figure 5.9.

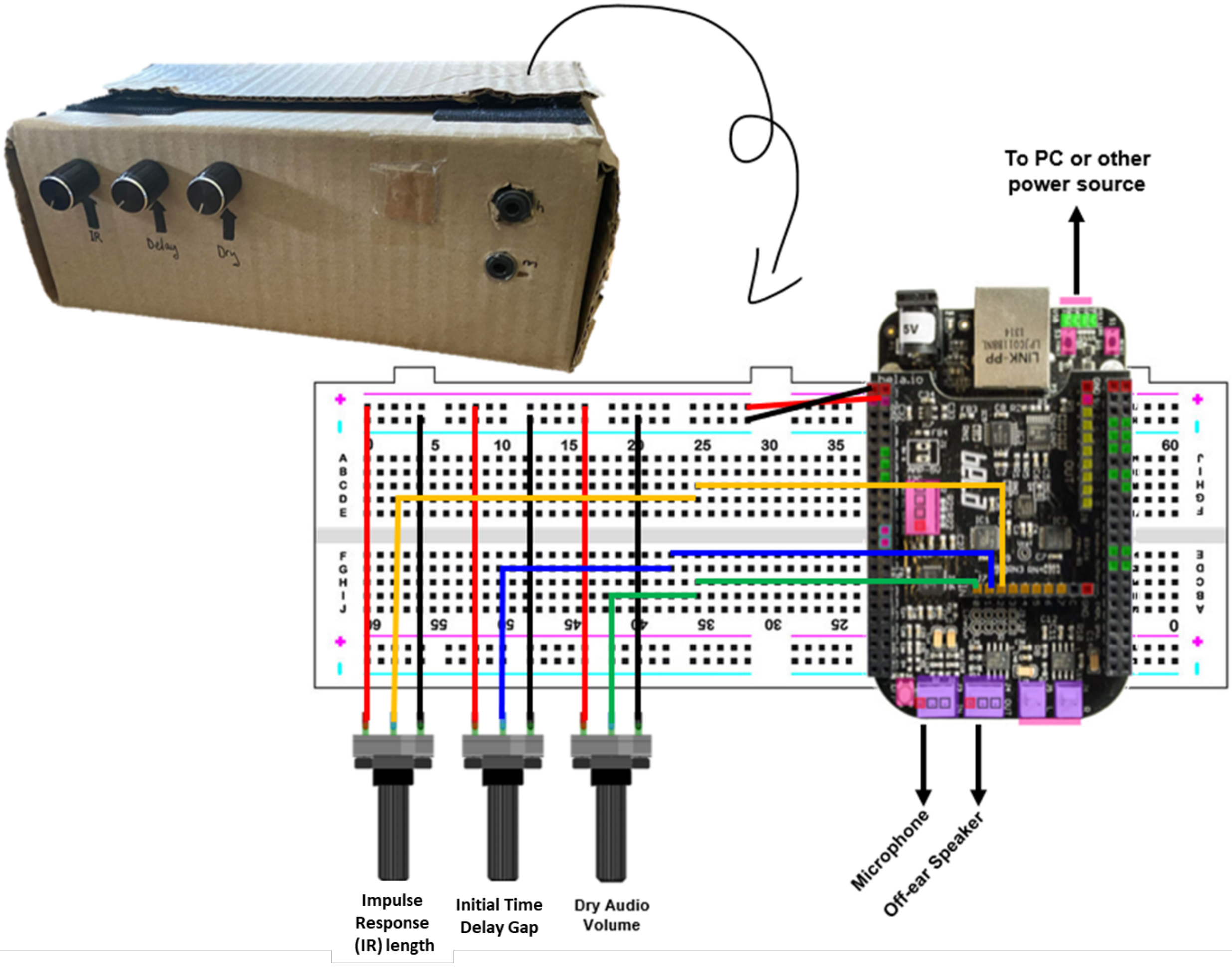


Figure 5.9: A photo of the Bela Box alongside a circuit diagram of the internal hardware.

An omnidirectional microphone was attached to the bottom of the headset in a permanent place, so that it would be in the same position for each user. The off-ear headphones from the deluxe audio strap attachment for the HTC Vive were disconnected from the main set of cables interacting with the computer and instead were routed through the Bela Box. It is these headphones that provide the real-time audio output for the participant.

The omnidirectional microphone takes in the live voice of the participant, which is then fed into the Bela board. In real time, the singer can manipulate the way their own voice sounds as it returns to them as though they are in a reverberant space, predominantly through altering the length of the impulse response and the ITDG. The headphones are off-ear, meaning that participants may not feel the need to include direct audio back out of the headphones. However, despite sitting off-ear, they do slightly alter the way a voice sounds back to the user, which is why the dry audio dial was included. As this device is intended to be implemented for a singing experience, where the user will be facing forwards predominantly, no head-tracking functionality was added to the Bela Box. As will be highlighted in Section 6.8.4, the Bela Board can be paired with head tracking sensors, but this was beyond the scope of this project [147].

5.6 Conclusion

The design and implementation of a novel, real-time, live-audio convolution system to allow a performer to alter acoustic variables of an impulse response as it is convolved with their own voice in real-time was showcased in this chapter. The Bela Box is a device which utilises the ultra-low latency capabilities of a Bela Board combined with a hybrid convolution method.

Created is a compact device that can convolve multiple RIRs, which can be switched between in real time alongside altering the ITDG, with a live audio input. The ability to add dry audio was also included, allowing the box to be used with in-ear and on-ear headphones if required.

Chapter 6

Some Sound Cues Contribute More to a Person's Perception of a Space than Others

> *"Suave locus voci resonat conclusus"*
>
> *(How sweetly the enclosed space responds to the voice)*
>
> - Horace, *Satires I, iv, 76* [207]

6.1 Introduction

This chapter describes the development of an experiment to test how relevant different sound cues are in a reverberant space when recreating the acoustics of that space. This is done through a combination of real-time convolutions, a VR headset, and a 25-part speaker rig organised in a 24.1 layout. Participants sing along with a recording of a real choir from a position within that choir, hearing their own voice back as if they were singing in the location of the original recording. The participant's live singing voice will be convolved with multiple IRs in turn, producing an "authentic" recreation of how a space responds to the voice.

This chapter first describes a pre-experiment undertaken to explore the concepts and techniques to be implemented in the main study. The motivation of the main experiment is explained in Section 6.3.1, followed by a system overview of the video and audio components. The recording processes for capturing the visuals and audio for the choir are detailed, followed by sections describing the rendering and installation required for the study. Then, the results of this project are recorded and analysed, followed by a discussion of the data, and potential future changes to the system. Finally, the chapter concludes with a summary of the project and the results.

6.2 We will rock you

Earlier in this research, exploration of experiments regarding performance in virtual, real spaces was undertaken, with initial inspiration coming from the research conducted by the University of York with their Virtual Singing Studio [143]. To become familiar with the setups and software required for this kind of research, a small volunteer choir comprised of a mix of choral singers and non-singers was formed. The intention was to create a short, interactive experience that could be built upon to explore further aspects surrounding virtual performance spaces. The choir learnt and performed the song *We Will Rock You* by Queen. This song was chosen as it has a catchy, repetitive chorus that is easy to pick up, whether or not you have heard the song before, due to its simple descending melody.

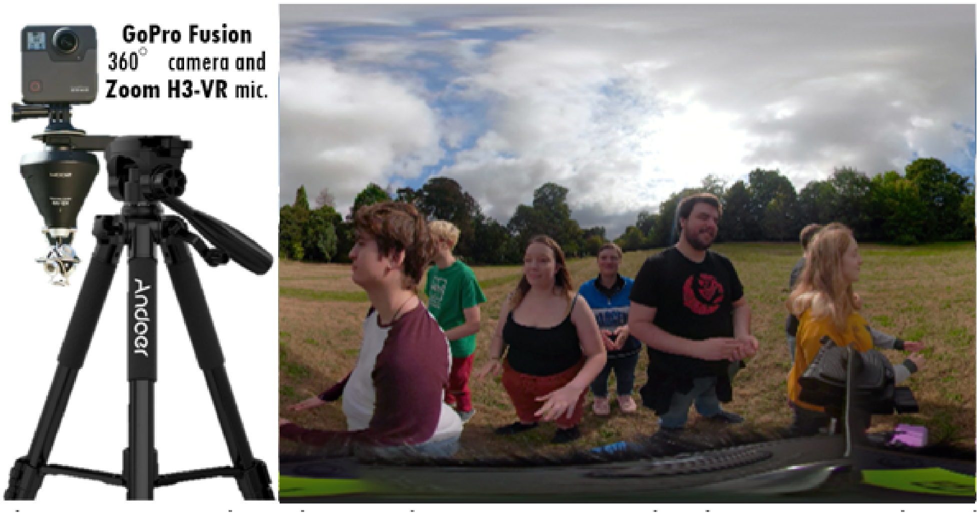


Figure 6.1: The rear 180° view from behind the virtual singer of the volunteer choir singing We Will Rock You, alongside the recording equipment setup used.

Figure 6.1 shows a snapshot from the rear view of a virtual singer within the choir, as well as the camera and microphone set-up, a GoPro Fusion 360° camera and Zoom H3-VR ambisonic microphone respectively [54] [80]. This experience was recorded using multichannel audio recording methods, specifically ambisonic audio, suitable for VR as well as 360° video capture techniques. The camera lens was set to a height of 170 cm with the microphone dangling a short distance below as shown in Figure 6.1. This camera-microphone pair are the eyes and ears of the virtual choir member. Once exported and stitched, the singer will be able to see and hear everything from that position. Therefore, it was decided that it would be more important for the camera to be at eye level than the audio be at ear level, otherwise the experience could be disconcerting for the virtual choir member, as discussed in previous work [120]. The piece was performed outdoors in an open space which presented various challenges for audiovisual data collection. There is a main road within earshot, and planes frequently fly over the area. This happens with some regularity, therefore recordings were timed to occur between flights semi-reliably. The ground is also not level, and based on previous weather, could be soft and unstable.

The choir were put into pairs or threes and given one verse each, with the intention of the virtual member joining in along with the chorus. The set-up of the choir can be seen in Figure 6.2, where the arrow signifies which way the front is for the camera and microphone.

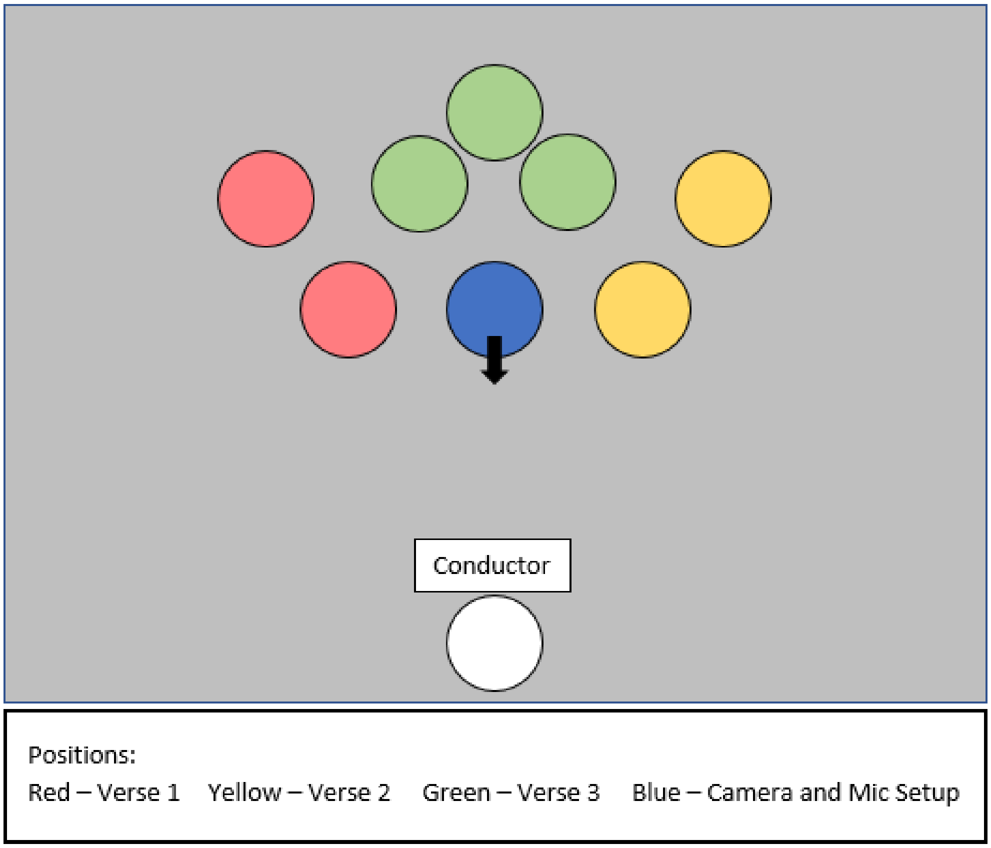


Figure 6.2: The positioning of the choir members and the recording set-up. The arrow represents the forward-facing orientation of the camera and microphone.

Each group was spaced around the camera set-up with one pair to the left, a group behind, and a pair to the right. This was done to enable to user to explore the directional audio of the setup, as each verse was passed around the group.

With this experience being an outdoor one, it was decided that no real-time convolution to simulate the acoustics for the outdoor environment was required. The absence of reflective surfaces encompassing the singers creates a negligible acoustic, hence simulation of this was deemed unnecessary. This exercise was helpful in demonstrating recording techniques both visually and audially for VR as well as for exploring performance in virtual environments.

6.3 Singing in virtual performance spaces

6.3.1 Motivation

From the above exercise, a final study was devised: when presented with live recordings from within a choir that you can sing along with, can participants decipher the correct acoustic properties for the space through singing themselves? Do the correct answers matter if the participant feels as though they have represented the space's acoustics authentically for themselves? Does this challenge require visuals to inform a participant of what to expect from the space, or does audio alone provide enough? As was investigated in Chapter 4, perception of the sound of a space through its visuals can cause a difference.

Virtual reality gifts us the opportunity to not only transport and transform places right in front of our eyes, places that we may never get to see in person for yourself for whatever reason, but also alter what is real. It enables us to push the limits of what feels natural or authentic to truly engage someone with an imitation of life. Reactive audio is the key for this.

The visuals in VR can only go so far, and current headsets have excellent visual controls to try to make someone feel physically present in a virtual space. With the onset of controllers that have separate finger movement, pressure sensors, and the ability to make virtual objects feel heavy in the real world, physical touch has grown leaps and bounds for VR, continuing to push for increased immersion. Whether you are virtually exploring a historical or fantasy building, would the ability to hear yourself back as though you were physically in the space make the experience more realistic?

The aim of this study is to create a virtual reality system, in tandem with a 25-piece speaker rig, that allows the user to replace a singer within a choir performance in a reverberant space. However, the exact audio properties of this space are to be determined by the participant in real time. Wearing an HTC Vive headset, a head-worn self-powered omnidirectional microphone, and by using Bela Box, the user should experience the venue of the original choir performance from the viewpoint of a member of the choir, and be able to hear themselves within that space as they sing along in real-time with a recorded performance [208]. The Bela Box would allow the user to alter which impulse response is being convolved, the length of the initial time delay gap (ITDG), and the volume of their direct (dry) audio as to best imitate what the singer believes to be the correct settings for the choral environment.

6.3.2 System overview

The construction of this study centres around two parts:

Audio

The Bela Box, as described in Chapter 5, is used to convolve the real-time audio input from the participant with one of five impulse responses with varying reverberant tail lengths. This is played back to the singer through the off-ear speakers of the HTC Vive with a deluxe audio strap. The choir audio recordings are output to a spherical array of 25 loudspeakers. A participant can stand in the centre of this speaker array, and through the use of an omnidirectional microphone attached to the VR headset, the user can hear themselves as though they are in the location of the choir as part of the performance.

Video

Wearing an HTC Vive, the user is able to freely visually experience singing with a choir from a position within the choir. This was achieved through recording 360° videos in multiple performance positions using a GoPro Fusion 360° camera.

6.4 Recording

6.4.1 Audio capture

The Zoom H3-VR microphone was used to record 4-channel ambisonic audio in Ambisonic B AmbiX format. For this study, the 4 channels would be upmixed to a 3rd order ambisonic file, requiring an SN3D normalised recording, which an AmbiX recording is. This upmixing was necessary for the software decoding the ambisonic audio for the speaker set-up, described in Subsection 6.5.2.

6.4.2 360° Video capture

The Go Pro Fusion 360° camera was used to record the 360° video of the singers. The combined set-up of the visual and audio recording devices is the same as in Figure 6.1, with the camera set to the eye level of the surrounding choir members, with the microphone dangling below. This decision was made as, as will be seen in Section 6.4.3, the choir performs on a staircase with shallow steps. Therefore, the height of the recording equipment changed slightly for each different recording location, but was always matched to fit the heights and spacing of the surrounding singers.

6.4.3 Song choices and recording layouts

The Jane Holloway Choir at Royal Holloway very kindly volunteered to be recorded for this project. They were asked to pick a piece of music that they knew and liked that was split into SATB (Soprano, Alto, Tenor and Bass) arrangement that they would normally sing in a reverberant space. The piece chosen was a choral arrangement of *Hymn to the Fallen* by John Williams, of which a 1 minute 15 second excerpt was recorded, as this was a piece familiar to them all. For each of the three recordings taken within the choir, the microphone-camera set-up was placed approximately at the centre of three sections of the choir, giving multiple perspectives on the experience based on which choir section a person could sing in. Just three recordings were made as on the day, only a small number of tenors and basses volunteered to be recorded, so one combined recording was made for them. The arrangement of the piece had multiple soprano and alto parts, so the recordings were made at the intersection of the two parts for each section, i.e., Recording 1 was made at the Soprano 1/Soprano 2 boundary. These capture positions are illustrated in Figure 6.3 and were made over three successive recordings.

A fourth recording was made so that the study could be run with participants who were not part of the choir normally. A simple, unharmonised recording of *Silent Night* was improvised on the day. This piece was chosen as it was suitable for all voice ranges when sung in unison at an octave of each person's choice, and is a piece that the average person probably knows the words and tune for. For this piece, choir members were asked to group together and stand next to somebody they would not normally stand next to ordinarily when split into sections. The recording set-up was placed in the centre of the choir, who were centred about the alter and aisle, with the forward-facing orientation of the camera

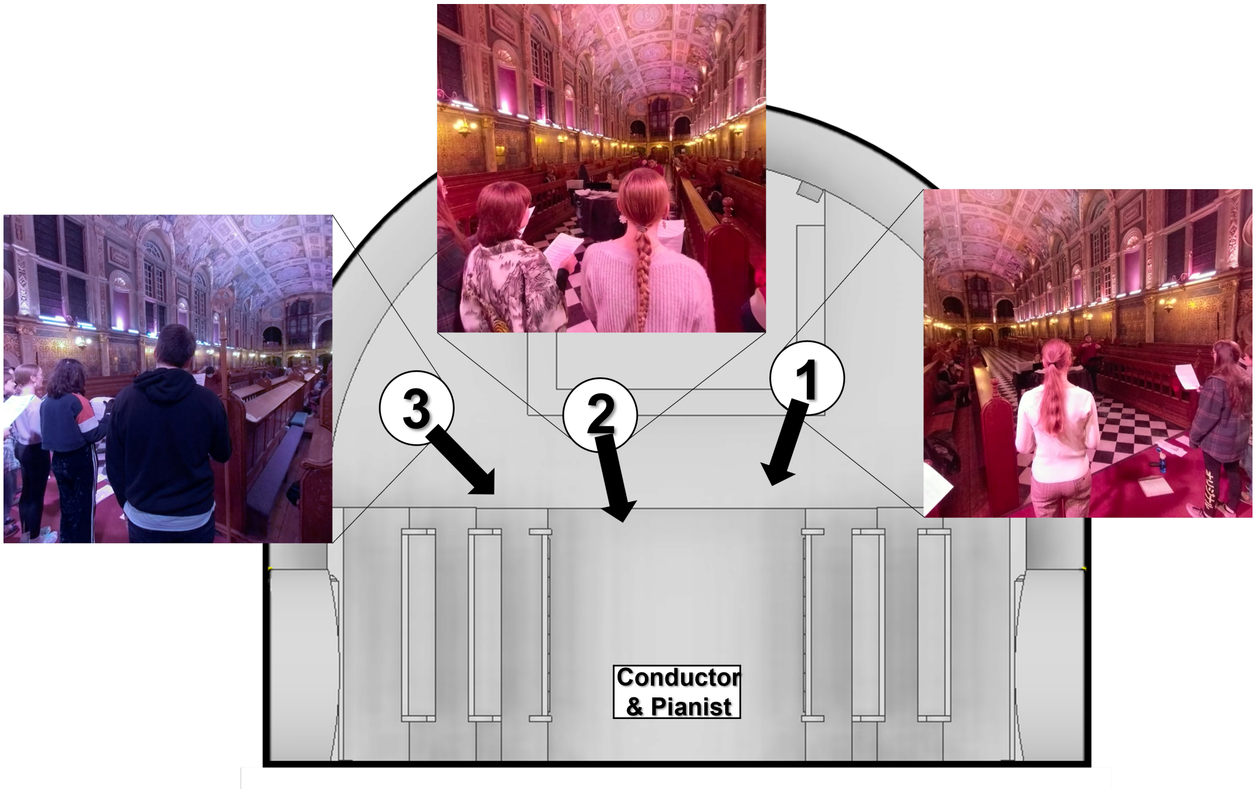


Figure 6.3: The three capture positions during the recording of the song *Hymn to the Fallen*. The white, numbered circles represent the camera locations, the arrows the forward-facing orientation of the recording setup, and the rectangular box the conductor and pianist. Recording 1: soprano point of view. Recording 2: alto point of view. Recording 3: tenor and bass point of view. The accompanying photos show the forward-facing view from the camera.

once again focused on the conductor. Figure 6.4 shows the front facing view of the recording set-up, and the rough position of the choir. Unlike the other piece, this song was sung a cappella as it was improvised.

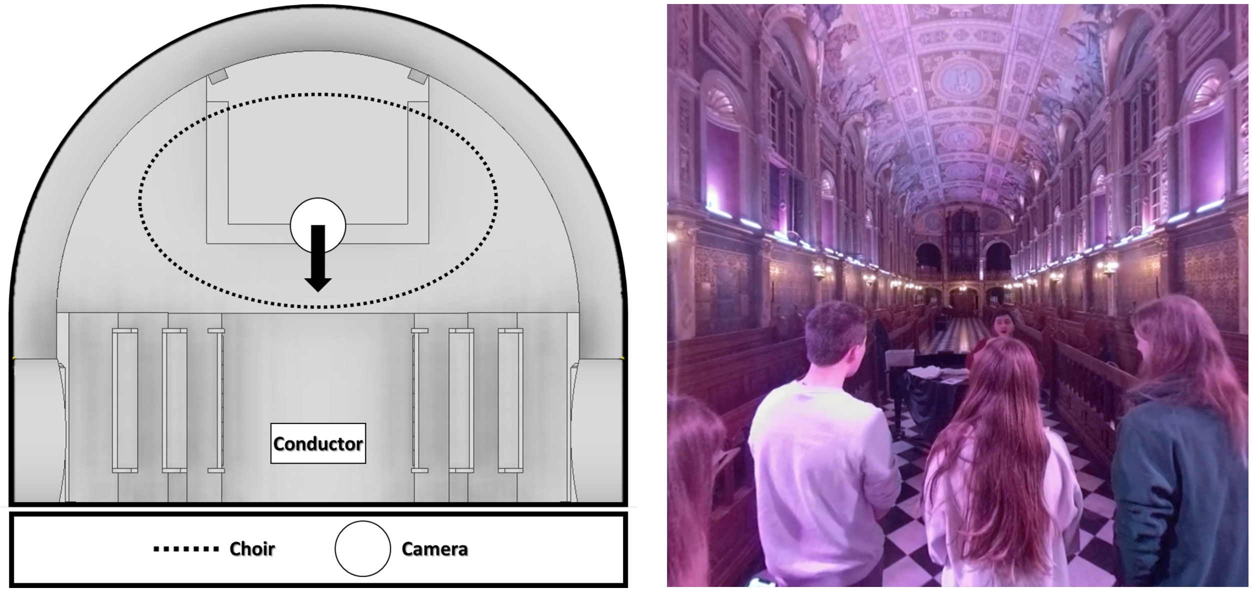


Figure 6.4: The recording layout for the improvised group piece. The arrow represents the forward facing direction of the recording equipment and the view of this is shown in the accompanying photo.

6.5 Rendering

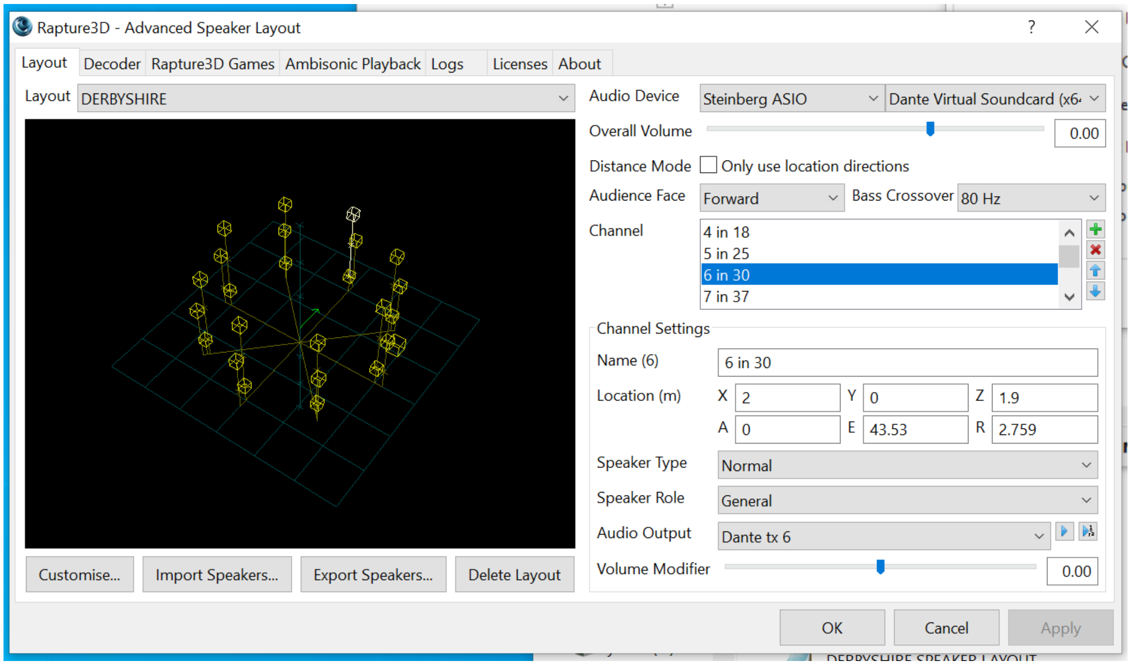
As already mentioned, there is a necessity for two separate systems to run different parts of the audio; the live feedback interactive side with the Bela Box, and the ambisonic audio of the choir from the recordings themselves. As the audio input side of the study is using off-ear headphones, a different kind of speaker setup was used for the choir recording; a 24.1 piece loudspeaker rig.

6.5.1 Adobe Premiere Pro

Adobe Premiere Pro was used to align the 360° video and audio files in the correct orientation. These videos were exported from the programme with no sound as the accompanying audio was to be played from the speaker rig. This step was still important as it ensured a mutual forward direction for all visual and audio components. Once oriented correctly, the 1st order ambisonic audio files were exported from Premiere Pro into REAPER.

6.5.2 REAPER and upmixing for loudspeaker setups

The 24.1 piece speaker setup was configured through Rapture3D from Blue Ripple Sound, with their advanced speaker layout tool. Figure 6.5 displays the working window with the specific speaker arrangement.



Layout: DERBYSHIRE

Audio Device: Steinberg ASIO | Dante Virtual Soundcard (x6·

Overall Volume: 0.00

Distance Mode: □ Only use location directions

Audience Face: Forward Bass Crossover: 80 Hz

Channel: 4 in 18, 5 in 25, 6 in 30, 7 in 37

Channel Settings

Name (6): 6 in 30

Location (m): X 2 Y 0 Z 1.9 A 0 E 43.53 R 2.759

Speaker Type: Normal

Speaker Role: General

Audio Output: Dante tx 6

Volume Modifier: 0.00

Figure 6.5: The working window for the speaker layout controlled by Rapture3D software.

Using the provided “Rapture 3D Parallel Decoder” in REAPER, 3rd order ambisonic audio files could be decoded to play over the speaker setup. The original recordings were only 1st order ambisonic files, so needed to be upmixed. Ambisonic audio is special because, as discussed in Section 2.3.4, the recorded soundfield is broken down into orthogonal functions instead of specific channels. Therefore, before decoding for the speaker layout, the audio files can be upmixed to 3rd order. This was achieved through the Upmixer from the COMPASS suite plug-ins [209].

COMPASS is a novel signal-dependent method for synthesis of ambisonic sound scenes using a more generic acoustic model than previous concepts. The Upmixer plug-in utilises COMPASS to take a lower-order ambisonic recording and upmix it to a higher-order ambisonic recording without inducing any unwanted artifacts into the system [210]. COMPASS also has an ambisonic decoder for arbitrary loudspeaker setups, however as Rapture3D was already implemented in the set-up, that was the software used in this study.

6.5.3 Bela Box

As described in Chapter 5, the Bela Box can control three acoustic properties of a participant's real-time audio input; the IR length, the ITDG, and the volume of the dry audio input. The latter two settings are changed through continuous scaling using their associated potentiometers. The dry volume is scaled from 0.000 to 1.000 through a conversion from decibels to linear amplitude. The ITDG is scaled from 0.00 ms to 0.20 ms after the live input. The IR lengths are set in advance by uploading separate files onto the Bela board which are then convolved in real-time. Each of the five IRs was mapped to a fifth of the available input range of the potentiometer, so by turning the dial to just over one fifth of the way around, the IR would change.

The impulse response lengths were chosen to have roughly the same percentage difference between them and were made with the original Pd patch shown in Figure 5.1. The IR lengths were named for their percentage length increase or decrease when compared to the original (100% IR) as 83% IR, 100% IR, 117% IR, 136% IR, and 154% IR. These were uploaded onto the board in a random order instead of from least to most reverberant. This was done for two reasons. Firstly, participants will be wearing a VR headset while using the Bela Box, so will not be able to physically see the dials. IRs with audible contrast were put next to each other to make it clear to the user when they had changed to a new one. Secondly, had the IRs been in order, there was a concern that participants would assume the real-world equivalent IR to be one of the middle three IRs on the dial, which could alter the way they answer the study. Therefore, the order from left-most to right-most dial position is 136% IR, 83%IR, 100% IR, 154% IR, and 117% IR. The full Bela Box code can be found on GitHub [206].

6.6 Installation and implementation

Multiple devices were required to be working in tandem to make this study possible. A diagram of the layout of the two rooms and the location of the devices within can be seen in Figure 6.6.

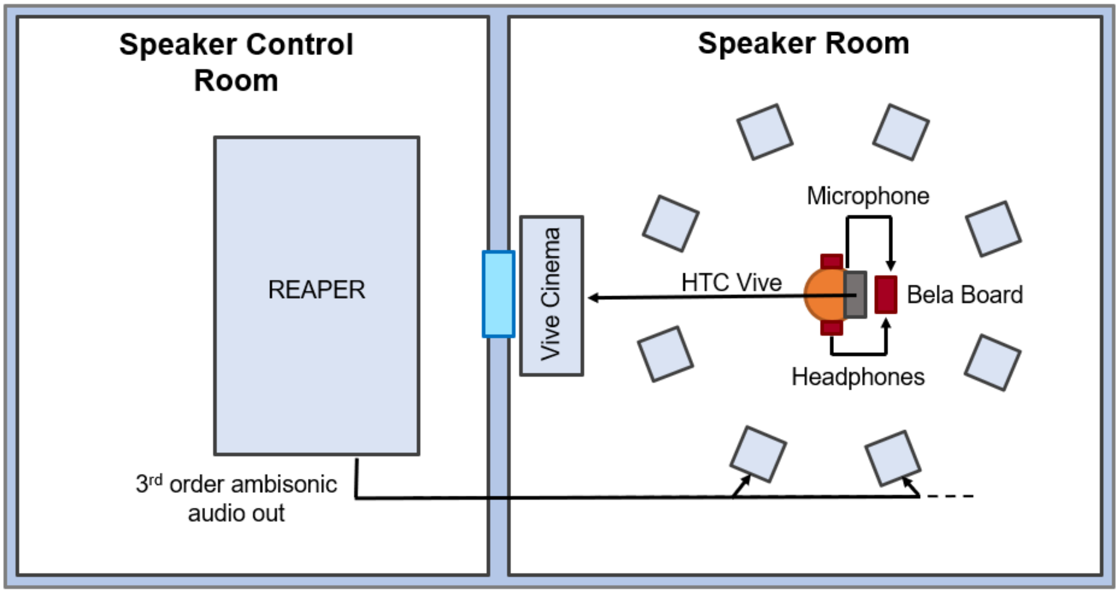


Figure 6.6: A diagram detailing the positions of all equipment used for the study across two rooms.

6.6.1 VR visuals setup

An HTC Vive headset with deluxe audio strap providing off-ear headphones, and a base station for tracking movement, were installed inside the speaker room on a separate computer. A monitor was set up at the window to the speaker control room to display the same view that the participant could see from within the headset. Despite not requiring this part of the setup to run any of the audio for the study, Vive Cinema was used to play the 360° videos as its user interface is friendly and easy to use, and it reliably plays these videos well (see Chapter 3). Photos of the layout can be seen in Figure 6.7.

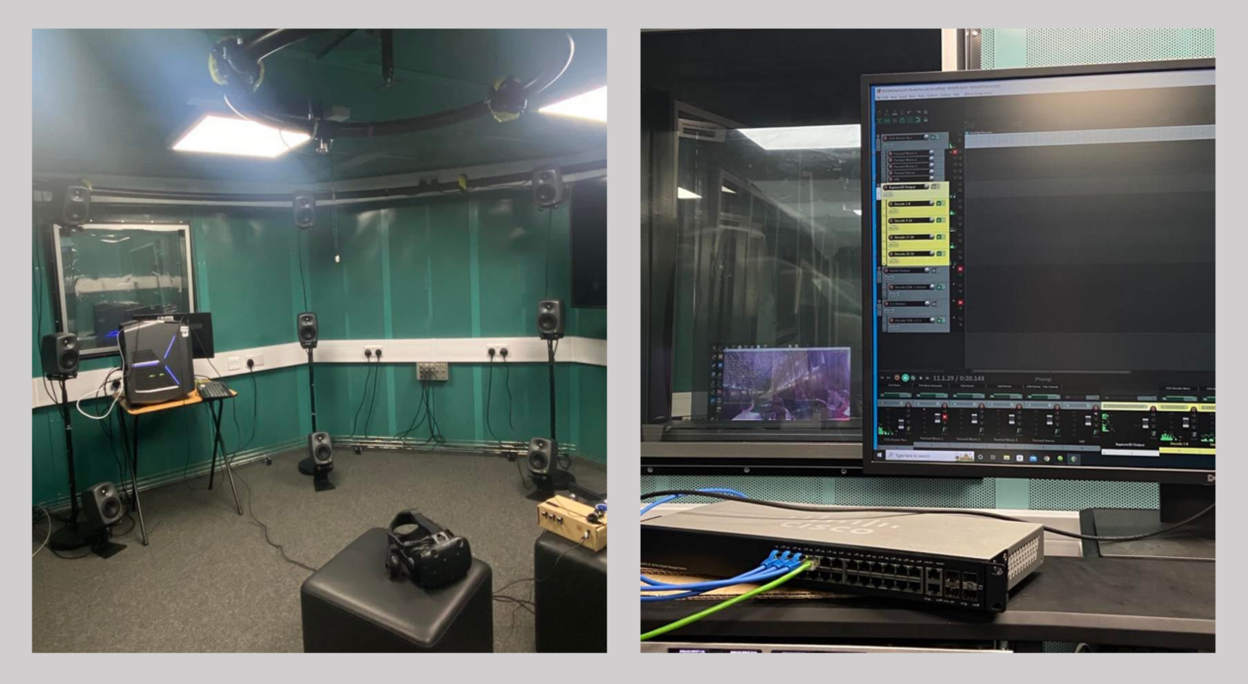


Figure 6.7: Photos detailing the layout of the experiment space.

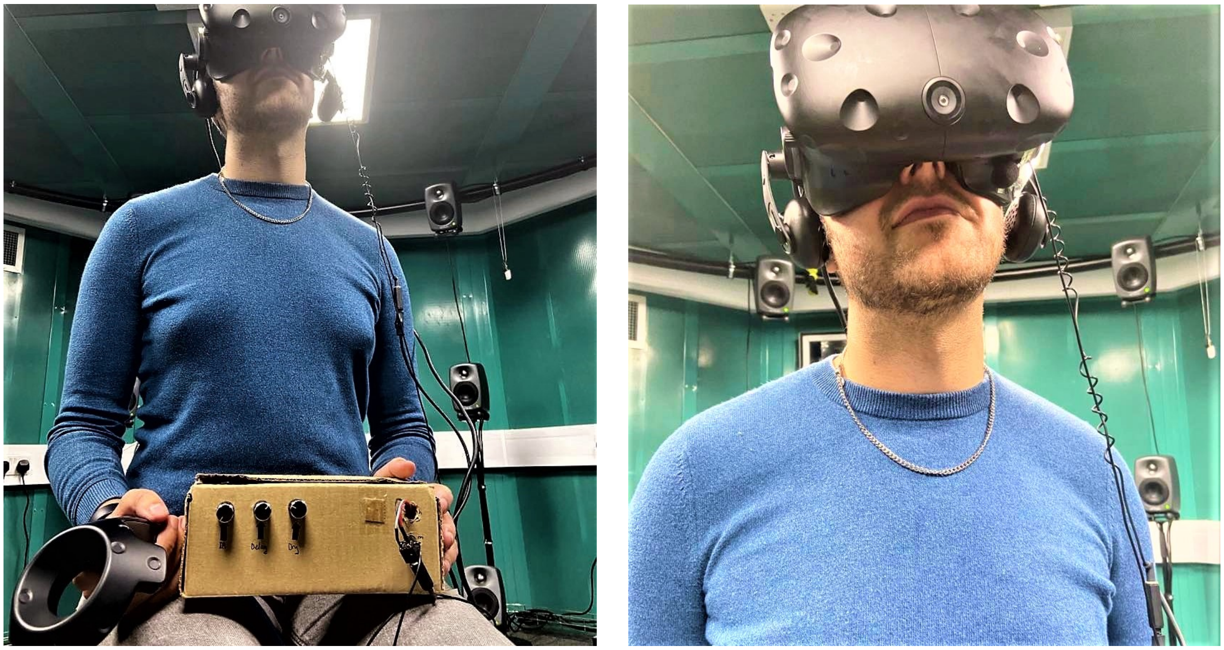
6.6.2 Choir audio setup

The upmixed 3rd order ambisonic audio files of the choir were loaded into REAPER in the speaker control room and, utilising the *Blue Ripple Sound Rapture3D decoder*, were decoded for the 24.1 speaker rig in the speaker room. As control of the choir audio and choir video are independent of each other, the 360° videos have a visible countdown at the start which will play on the monitor facing the speaker control room. This allows for the audio and video to be synced up despite this process being manual.

6.6.3 Real-time audio setup

The Bela Box was installed in the Speaker Room by connecting the off-ear headphones of the VR headset to the Box, and plugging in an omnidirectional microphone that was taped in a fixed position on the base of the headset. It is through these re-routed headphones and this microphone that the real-time audio component of this study will be provided. Figure 6.8 shows the Bela Box, headset, and attached microphone on a participant. The box was turned around to show the dials. During the study, the dials face the user.

As described in Chapter 5, the microphone takes in the live voice of the participant, which is then convolved in real time with the Bela Box. The singer can manipulate this convolution with the dials by setting the length of the impulse response to one of five set options, and by changing the ITDG. The headphones are off-ear, however they do slightly alter the way a singer's direct audio sounds back to them, which is why the third dial which can add dry audio to the mix was also included. It is hoped that the off-ear headphones combined with the 24.1 piece surround speaker setup lead to a natural integration of the two audio components for the singer.



(a) (b)

Figure 6.8: Photos showing the positions of the off ear speakers and microphone and how they are connected to the headset and Bela Box.

6.6.4 Procedure

Participants were given an instructions document to read through which explained the core purpose and aims of the study before being verbally instructed with further details. In this study, participants were directed to sing along with one of two songs as they virtually joined the Jane Holloway Choir singing in the Royal Holloway Chapel. They were then presented with the Bela Box, shown in Figure 5.9, and were told how the box could alter their voice in real time, as well as explaining where different audio sources (i.e., the choir and their own voice) would be emitted from. Participants were told that they would be given three chances to sing along with a piece of music. For the first 2 run-throughs and in the breaks in between, they were encouraged to alter the way their own voice sounds freely, with the intention of imitating how they thought they would sound in a choir in the Chapel. For the final sing through, they were told to settle on the settings they believed were the most natural for the Chapel and try to immerse themselves in the experience, as the questions to fill in at the end would all be to do with their thoughts and perceptions while singing with the virtual choir.

Participants were then given ample time to play with the various audio settings alone in the space so that they could get a feel for the limits of how they could manipulate their voice without the potential of feeling inhibited by another person listening. Between each sing through of the piece of music, the researcher re-entered the room to check on the participant and see if any further questions had arisen. Once the participant had finished giving their feedback, a short piece of code is added back onto the board to print off the values set on the dials.

6.7 Results

6.7.1 Participants

23 participants took part in the study, providing written informed consent, in accordance with the Royal Holloway Research Ethics Committee. Participants were comprised of a mix of students from the choral choirs on campus, and casual or non-singers. They were not required to be familiar with virtual reality, or to sing regularly, to take part, however participants were asked about their previous experiences in both areas as part of the study. Neither gender nor age were considered for this study as it was deemed inappropriate.

The majority either currently, or had previously, participated in some form of regular group singing (15 of the 23 participants), however only 4 sang regularly at the time of the experiment. Not all of these singers had experience with the Royal Holloway Chapel however. 13 of the participants had some form of experience singing, speaking, or listening in the Chapel prior to the study, and 3 of this subset were people who had rarely or never sung in a group setting before. Most of the participants were also familiar with virtual reality, with 9 people having significant prior experience, 8 having had some, and 6 having never experienced VR before. All participants were given as much time as they felt they needed to get familiar with each aspect of the study regardless of their answers to the above questions.

6.7.2 Dial values

Participants could manipulate three audio properties to alter the way their voice sounded back to them; the IR applied to their voice, the time for the first reflection to return, and the volume of their dry signal.

As mentioned in Subsection 6.6.3, participants were informed that the dry signal slider was included purely for preference if they found the off-ear headphones to inhibit the ability to hear themselves back clearly. Just over two thirds of all participants did not feel like they required their direct audio to be played back in real-time, leaving the dial volume low enough to be inaudible. Three participants did choose to have the volume fully up, but none commented on it in their questionnaire, nor did it seem to affect their answers to the main questions posed by the form.

Five impulse responses with varying lengths of reverberant tail were loaded onto the Bela board to choose between, with participants being informed that one of them was the real IR for the chapel. They were also aware that the IRs were not arranged from least to most reverberant but were instead randomly ordered. The results for this dial can be seen in Figure 6.9. About one third of participants chose the correct reverberant tail length for the chapel, with 47.8% of the total participants choosing one of the two similar IRs (*83%* or *117%*).

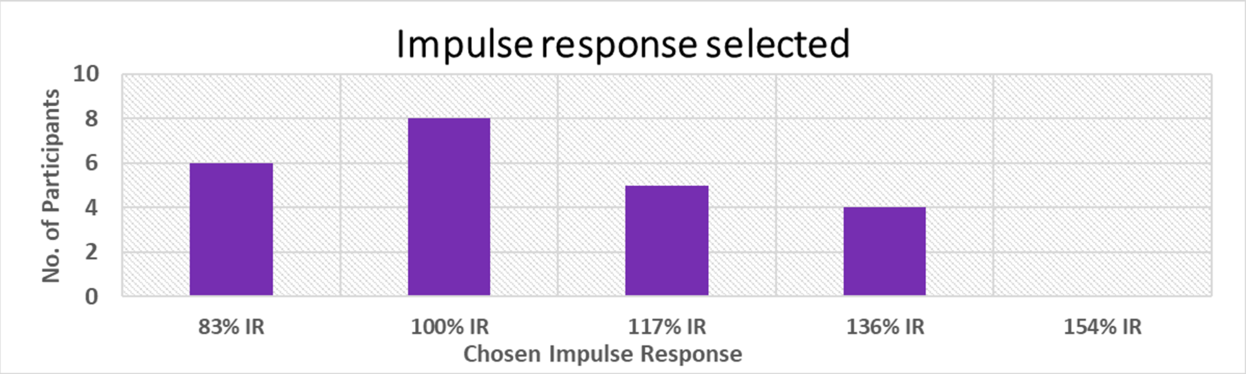


Figure 6.9: A graph showing the spread of impulse responses selected by participants. *100%IR* is the actual impulse response for the chapel.

The ITDG dial produced a varied spread of results, which can be seen in Table 6.1. The outermost values of the table were at the extremes of the available range participants could explore. Unless they asked, it was not disclosed to the participants whether or not their dial values were the correct ones for the chapel.

Table 6.1: A table highlighting the spread of results for participants choosing what they thought the correct ITDG for the first reflection for the chapel was.

Deviation from correct ITDG (ms)	-5	-3	-2	±0	+1	+2	+3	+5	+6	+9	+11	+14
No. of participants	7	2	1	2	2	1	1	1	1	1	1	3

6.7.3 Questionnaire results

Participants could take part in one of two modes of the study; the mixed voice *Silent Night* choir, or singing specifically with a part in an SATB arranged choir singing *A Hymn to the Fallen*. Those who took part in the latter had learnt the piece of music whilst singing with that choir, whereas anyone was welcome to sing along with the former. The questionnaire was presented to the participant after their third sing along, and each person was asked to reflect on the statements based on their final sing through, which was when they had settled on which vocal settings they thought were correct.

Silent Night **questionnaire results**

Participants who sang along to *Silent Night* were presented with a set of 6 statements and asked to rate them on a scale from one to five, the latter being the “positive” end of the responses:

(Q1): Ease of hearing your own voice (*difficult to easy*)

(Q2): Ease of blending in with the choir (*difficult to easy*)

(Q3): Ability to keep tempo (*difficult to easy*)

(Q4): Reverberation quality (*not natural to quite natural*)

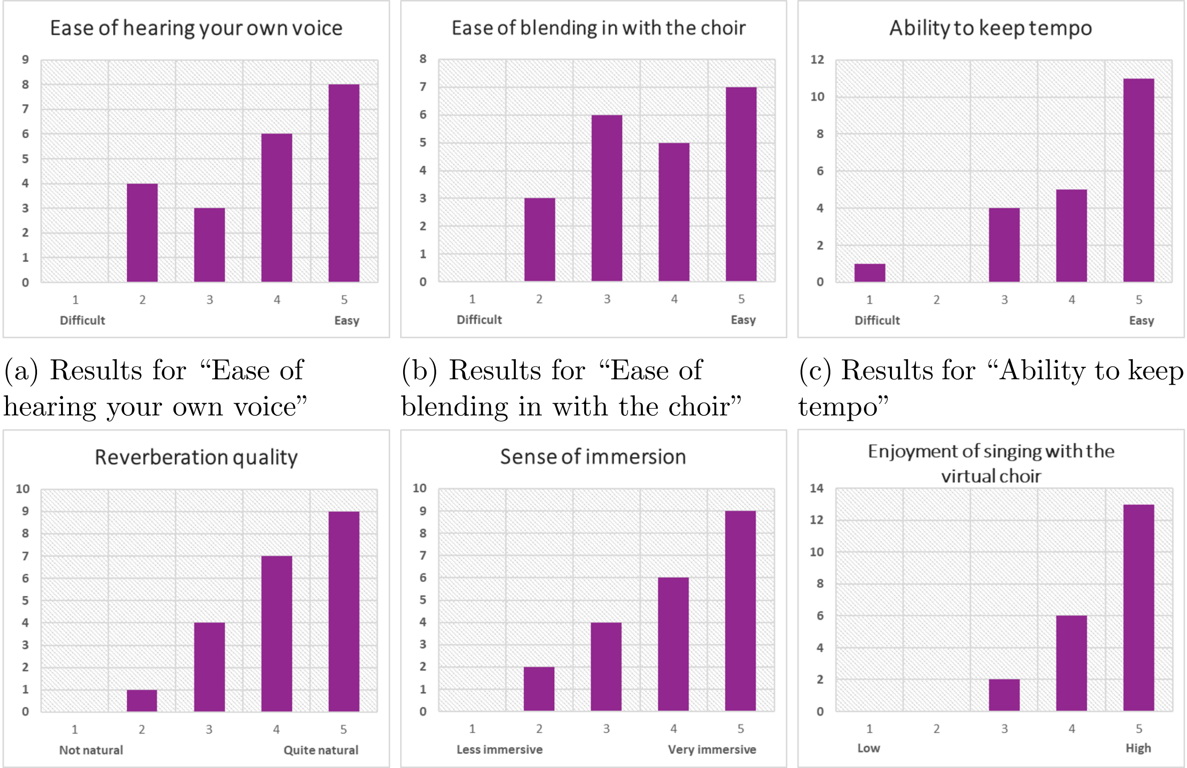
(Q5): Sense of immersion (*less immersive to very immersive*)

(Q6): Enjoyment of singing with the virtual choir (*low to high*)

Followed by three general questions about the experience:

1. What did you think of the experience as a whole?
2. Would you take part in an experience like this again?
3. If you have sung in the chapel before, did this experience feel natural/realistic? Please explain your response.

Figures 6.10(a) to (f) show the ratings given by participants for the 6 statements.



(a) Results for "Ease of hearing your own voice"

(b) Results for "Ease of blending in with the choir"

(c) Results for "Ability to keep tempo"

(d) Results for "Reverberation quality"

(e) Results for "Sense of immersion"

(f) Results for "Enjoyment of singing with the virtual choir"

Figure 6.10: A set of bar graphs displaying the answers for participants who sang with the mixed unison choir singing *Silent Night*.

When asked about how they felt about the experience as a whole, 17 of the 21 *Silent Night* singers responded positively, talking about how immersed they felt, and that it was “very fun”, “pretty cool”, and “very interesting”. Those that responded less positively still said that they had fun, however stated how they struggled to be immersed as they found it hard to blend their voice with that of the choir. Either due to the mismatch from “the sounds being produced from different speakers”, how it “felt wrong to be sat down”, or the fact that they were singing Christmas choral music and were not a fan.

When asked if the participant would take part in something like this again, 85.7% responded with a confident yes, and only one person with “probably not”. The few people that elaborated further talked about how they would want to see it used in further activities, perhaps exploring it as a “training mechanism to understand how different acoustics affect the music”, or just with more harmonies. One person went on to explain how they enjoyed it because “it [was] a way that I [could] take part in singing projects that I would usually avoid”, as they were not confident singing ordinarily.

Of the 21 people who sang *Silent Night*, 9 of them had sung in the chapel before, and one other had sung in other churches/chapels. Question 3 of the general questions asked participants who had sung in the chapel before if they felt that the experience felt natural/realistic to them. All those who responded did so mostly positively, with the general consensus that it did feel “pretty realistic”.

Those that had sung in choral groups before, including in the chapel, spoke highly of the experience, with only one of them commenting that although it was realistic, they felt it was “slightly unnatural”. Most stated how they felt that “the reverb in the chapel can often sound quite distant and the emulator replicated this very well;” that “yes it was weirdly real... The echo really made it feel like I was there;” and “it did really feel like I was in the chapel. I thought the effect was very natural.” Others who also had choral experience at the university talked about how the aural experience "...differed greatly from my expectations of how it would sound but ... [it] certainly sounded realistic to a chapel."

From those who had sung in the chapel or another church environment before but not in a regular group singing activity, most talk about how "eerily accurate" some of the "echoes on the first dial" were, with a single person once again commenting on an unnatural feeling, stating how hearing "...my voice played back sounded a bit like it was digitally recorded and played back (which it was)." The person who had not sung in the chapel before commented on how they "got the real feeling that there was a mass space around you rather than being a small room".

Hymn to the Fallen questionnaire results

Three members of the Jane Holloway Choir (the choir in the recordings used) were available to take part in the other mode of the study; one that looked more specifically at those who often sang in the chapel in a group setting and in SATB form. In this version of the study, participants sang with either the sopranos, altos, or tenors and basses to an excerpt of *A Hymn to the Fallen*. One singer from each section was available.

These participants were asked two extra questions (Q4 and Q8) compared to those who sang *Silent Night*:

(Q1): Ease of hearing your own voice (*difficult to easy*)

(Q2): Ease of blending in with the choir (*difficult to easy*)

(Q3): Ability to keep tempo (*difficult to easy*)

(Q4): Ease of keeping in tune with the choir (*difficult to easy*)

(Q5): Reverberation quality (*not natural to quite natural*)

(Q6): Sense of immersion (*less immersive to very immersive*)

(Q7): Enjoyment of singing with the virtual choir (*low to high*)

(Q8): Similarity to singing with the choir in real life (*low to high*).

Responses to these statements for the three singers can be seen in Figure 6.11.

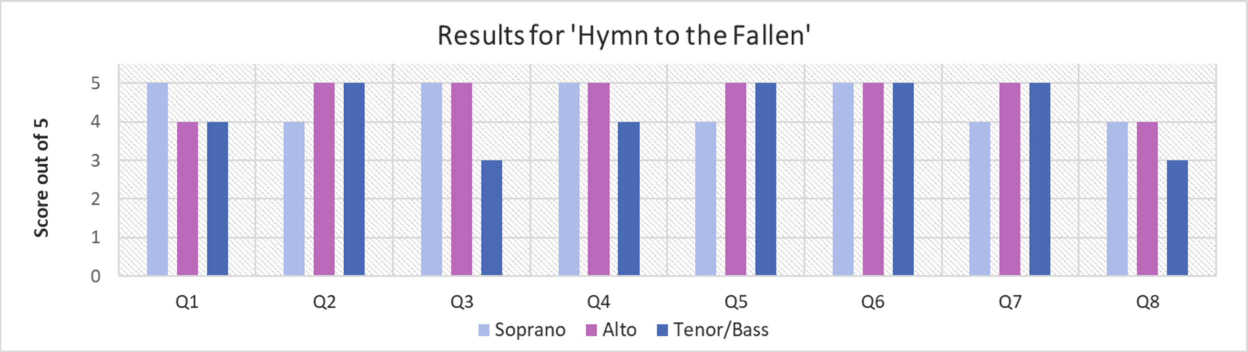


Figure 6.11: A graph showing the responses from members of the Jane Holloway Choir to the statements posed.

The participants were then asked the same set of general questions as listed in the *Silent Night* questionnaire results. All three enjoyed the experience and would take part in something similar again, saying how "singing with the virtual choir was quite realistic to singing with the live choir" and how it was "...interesting to be part of something that I am familiar with in real life virtually. I [thought] it really did make you feel like you were in the choir and with the sopranos sounding louder because I was near them it just felt quite real."

The singers had sung very frequently in the chapel before, and were also asked whether they thought the experience felt "natural/realistic". All three stated that they thought it did feel real and natural: feeling that they were able to "follow the conductor's tempo and blend with the voices easily" as well as commenting on the reverberations feeling both "real" and "full".

6.8 Discussion

6.8.1 Discussion: dial values

None of the participants correctly identified both the correct reverberant tail length and ITDG exactly, however taking the exact value of the delay is not very fair. A different classification of what will be described as correct will be explained below in this section.

Reverberant tail length

As stated in Section 6.7.2, the five impulse responses were loaded onto the Bela board in a random order. This was for two main reasons; to stop participants from attempting to guess which one was correct based on its position in the order, and to increase the difference between subsequent IRs to improve user clarity due to their vision being obstructed while in VR.

Participants who had familiarity with the chapel chose the correct impulse response in 38.5% of cases. Comparatively, 30.0% of respondents who had no familiarity with the chapel chose the correct IR. This would suggest that familiarity with the space has a minor impact on the ability to ascertain an accurate soundscape. Another characteristic is shared between the respondents that correctly identified the most accurate IR however: experience of group choral singing in Royal Holloway's chapel. Of the 6 participants with group choral experience in the Chapel, 5 correctly chose the most accurate IR. This includes 4 participants who mentioned singing with the Jane Holloway Choir this academic year. Therefore, those who had regular exposure to group singing in the Chapel had the highest likelihood of identifying the real impulse response for the space.

Initial time delay gap (ITDG)

It has been shown that even for singing, when compared to instruments for example, where you can also hear yourself internally through bone conduction, the sensitivity for which a person can perceive a difference in latency of sound falls anywhere between 3ms and 10ms for an in-ear speaker, which only increases for external speakers [211].

Therefore, it would not be remiss to consider a broader range around the measured ITDG when considering which participants were correct in their choice. Table 6.2 arranges respondents' choices into ±5ms ranges extending out from the recorded value.

Table 6.2: A table displaying respondents' choices of ITDG grouped into ±5 ms ranges.

Deviation from correct delay (*ms*)	within ±0	within ±5	within ±10	within ±15
No. of participants (out of 23)	2	17	19	23

When treated this way, 73.9% of participants chose an ITDG value within 5ms of the measured time, with only 17.4% falling outside of a ±10ms range. Participants familiar with the Chapel were within ±5ms of the ITDG in 69.2% of cases, compared to 80.0% of those with no prior experience, once again showing that familiarity with the real environment has a minor influence on the ability to discern the correct delay gap.

Similarly to the reverberant tail length results, experience of group choral singing in the real chapel did influence a participant's ability to determine the ITDG correctly, with 5 of the 6 experienced singers once again choosing the correct value. This continues to make them the majority of the correct values for both dials for people who have prior familiarity with the Chapel.

When considering responses for both dials, including the ±5ms ITDG window, 26.1% of participants got both values correct. Of these subjects, two thirds of them come from the regular chapel singers, with the other third having had no experience with the chapel ever.

6.8.2 Discussion: questionnaire

Overall, the response to the study was overwhelmingly positive which can be clearly seen from both Figures 6.10 and 6.11, with the combined mean rating for each of the six main questions for both groups represented in Figure 6.12.

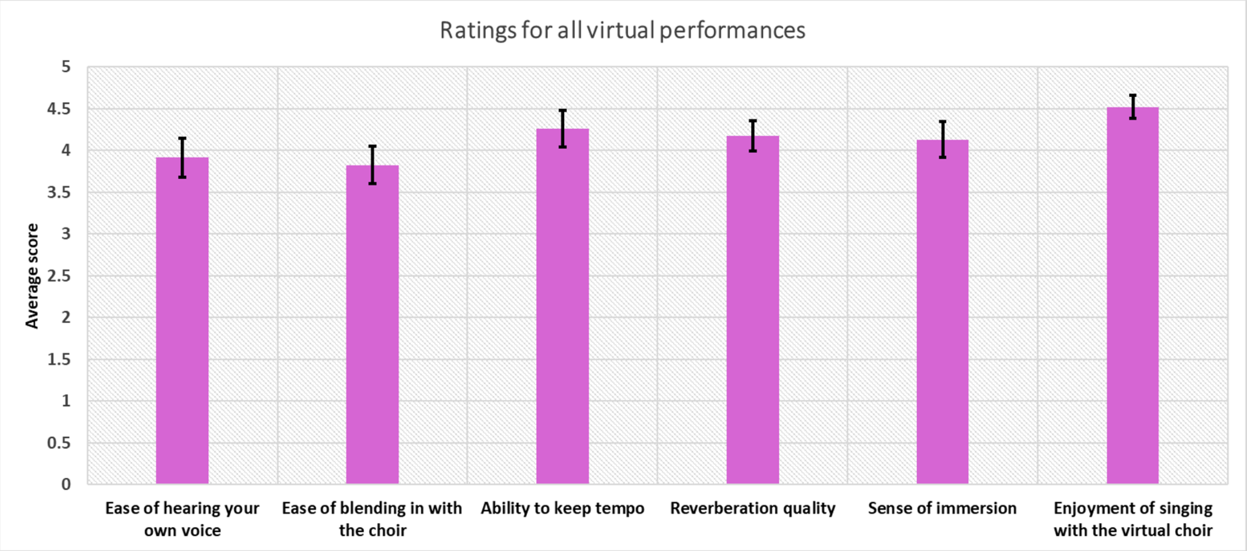


Figure 6.12: A graph showing the mean scores (and standard error) of singers' responses to the questionnaire

"Enjoyment of singing with the virtual choir" was the highest rated answer, however all averaged to close to 4 out of 5 and above, suggesting that even if people were not sure what the correct settings for their voice on the Bela box should be, they still felt immersed in the experience and enjoyed it. The standard error for each question never drops below 3.60, which further suggests that if this study were to be run again or installed permanently as an experience that it would be enjoyed.

For the three members of the Jane Holloway Choir who sang along to *Hymn to the Fallen*, the tenor/bass singer's average rating of the experience was lower than the others, being the only one to rate parts of the experience with a 3 out of 5 (see Figure 6.11). They mention in their written response how "singing with the virtual choir was quite realistic to singing with the live choir", despite their ratings. This participant did verbally mention something that may have influenced their answers: unlike the soprano and alto singer, the tenor/bass singer virtually stood directly behind themselves in the recorded videos. They brought up how disconcerting it was to be able to see the back of their head during the study, which may have had an influence on their experience as a whole, especially in response to Q8.

Which sound cues contribute most to a person's perception of space when recreating acoustics? Those who mentioned the reverberation and delay in their written responses brought up similar points. Of the 23 participants, 8 talked about the "echo" or "reverb" settings. Some of these respondents stated that it was "difficult to get the reverberation right" but once they had, "the echoes on the first dial really did make my voice sound like it did in the chapel in real life". It was obvious to most when the reverberation felt wrong, which is backed up by the results, as none of the participants selected the most reverberant IR, with the second least popular being the next most different from the real space.

The ITDG was not brought up often in the responses to the questionnaire, with only 3 people discussing it. Two of these participants state how they found the ITDG "really noticeable when wrong", although this did not make them confident in their choices. The third directly said that they found it hard to sing along in time because of the delay, and that they "maybe [I] just didn't find the right spot on the second dial."

With the majority of participants (73.9%) falling within a reasonable range of the ITDG, it could be argued that this ITDG value was the most important of the two main sound cues for participants to manipulate, and most settled in a range within which they may not have been able to hear a difference. In comparison, the results for the reverberant tail lengths were more varied. Although 82.6% of subjects chose either the correct, or one of the two similar, IRs, each one makes the users' voice sound noticeably different. Therefore, an acoustic recreation of a real-world space can probably get away with the IR not being completely correct as long as the latency of the system is properly accounted for.

6.8.3 Discussion: realism and immersion

One of the comments left by a frequent choral singer, although not with the choir recorded for this project, stated something in their results that really encapsulates what the research over the last four chapters has led to. When asked if the experience of singing in the study "felt natural/realistic" to them when compared to singing live in the Chapel, they responded that "It differed greatly from [my] expectations of how it would sound, but [I] am not sure whether that was [my] expectations or the simulation. It certainly sounded realistic to a chapel."

This participant has unknowingly described the grey area of what realism is when recreating a real-world environment. Returning to the initial definitions of what defines realism, authenticity and naturalness from Chapter 1, this comment neatly encapsulates the difficulty of assessing virtual recreations of real spaces, harkening back to earlier literature discussing hyper-realism. The singing experience was not what the participant expected; their singing voice did not sound like they imagined it would in the chapel. The singer is unsure if this is because they went in with an "incorrect" expectation, an expectation that was *authentic* to their perception of what the Chapel sounds like but not *realistic*, or if the simulation was not a *realistic* reflection of what the Chapel actually sounds like, therefore contrasting with their prior real-world interactions. The participant does however feel as though the experience was realistic to " *a* chapel", to a reverberant singing space. Just not necessarily *the* Chapel.

All other participants, bar one, who answered this question said that they felt that the experience was accurate to that of singing with a choir in the Chapel and that it did feel natural. This is reflected in the numerical results. The previous subsection demonstrates that participants were immersed within the singing experience, with an average score of over four out of five. From these results, it could be argued that the experience created by this study is, by the definitions adopted in this thesis, a realistic recreation of the acoustics of the Chapel.

6.8.4 Further exploration and improvements

Having more singers who regularly perform in the chapel take part in the study would have been interesting. The initial pattern that has emerged from this study suggests that generally, they are more capable of identifying the correct acoustic settings, however, this study was run with a small sample size of people, so to know if this is true, it would have been ideal to engage with more singers.

Secondly, as explored in *The Effects of Latency on Live Sound Monitoring* [211], different people have different abilities to discern latency, therefore perception of the ITDG would vary between participants. Running an extra part of the study to explore each participant's sensitivity to latency, rather than using a general window of 5ms from the true value, would have allowed the discussion of ITDG results to be tailored to each person's ability.

Some questions in the survey pertained to the participants' experience with how the virtual environment affected their ability to competently sing within the choir. The answers to questions concerning the ease with which the participant found staying in tempo and tune with the recording will likely be heavily influenced by the way the participant perceives their own skill in choral singing.

A participant that feels confident with choral singing will likely not struggle to maintain tempo and tune regardless of the virtual environment, but would still provide a high score in the survey on those questions. Therefore, if this study were to be run again in some form, it would be advantageous to have participants sing both physically and virtually with the choir and re-tailor the questions to be focused on the comparison between the two experiences.

Finally, although only used for managing real-time audio in this research, the Bela board can be utilised in a multitude of other immersive ways. In tandem with a head tracking sensor, Bela has been used to create an interactive binaural scene that is explored through the movement of your head [147]. It has also been developed for use with real-time diffusion across multiple speakers as a flexible tool for controlling spatialisation manually for multiple speaker arrays [212]. Furthermore, outside the world of audio, Bela boards have been used in the creation of interactive wearable artwork, where the audience experiences seismic activity [213]. To create something potentially more cohesive, or not reliant upon an immovable speaker rig, it would be interesting to adapt this project to work just with a headset and Bela board. For example, the members of the choir could be recorded separately in a studio or with spot mics during the recording, and later fix those sound sources in place, like with the project detailed in reference [147], creating a binaural scene that only requires headphones.

6.9 Summary and conclusions

This chapter outlined the process of using a Bela board, VR headset, and 24.1 piece speaker rig to allow a user to manipulate their singing voice in real time to perform alongside a virtual choir. The impulse response, and choir audio and video were recorded in a similar way to that of the previous chapters. Pure Data, and then ultimately C++, were implemented to create a "zero-latency" convolver which allowed a participant to change different attributes of their convolved voice before it returned to them.

The main impetus for this study was to ascertain whether certain audio cues were more or less important for immersion in an acoustically reverberant space than others. It was shown that for the reverberant tail length, 82.6% of subjects chose the correct, or a similar (within ±17%), impulse response, and 73.9% chose the correct ITDG within ±5ms. As described in Section 6.8.2, it can be argued that having the correct impulse response is the less important acoustic factor, as 65.2% of participants were happy with an incorrect reverberant tail, whereas 17% were happy with an incorrect ITDG.

Section 6.8 shows how those who had prior experience singing in a choral group in the Chapel were correct more often compared to those with little or no prior experience either being physically present in the chapel, or with singing in a choral group. As a whole, the experience was enjoyed, and singers were immersed in the virtual choir (see Figure 6.12), with 20 of the 23 participants stating with confidence that they would take part in a similar experience again. When looking at the responses to the question aimed at participants with prior singing experience in the Chapel, we can conclude that the experiment did provide a "realistic" recreation of the acoustic properties of the chapel.

Chapter 7

Conclusions

The main aim of this research was to investigate the use of virtual reality combined with immersive audio techniques to explore real locations, both indoor and outdoor, virtually through acoustic natural representation. In order to investigate this, three studies were created to explore current and new methods of utilising various auralisation techniques and software.

Participants demonstrated through these three experiments how virtual reality paired with these acoustic processes does allow us to naturally, although not necessarily identically, recreate these environments to such a degree that they feel natural. The three studies showed that, when compared to falsified or modelled environments, real acoustic environments cannot be correctly distinguished from modelled ones.

7.1 Summary of thesis

A summary of the work presented in this thesis is as follows.

Chapter 1 - Introduction

Chapter 1 introduced virtual reality and how immersive experiences have become more common and accessible over the last decade through the popularity of VR. This chapter also defined terms such as ‘immersion’, ‘realism’, and ‘authenticity’ for use in this body of work. The aims of the thesis were presented, followed by the hypothesis and how it would be tested. The novelties of the work were then highlighted, and the structure of the thesis outlined.

Chapter 2 - Theory and literature

Chapter 2 began by introducing core concepts in the field of virtual reality and immersive audio, starting with the rise of access to VR for the wider public, and the real-time 3D engines through which one can create content for these devices. It went on to explore and describe how sounds can be formatted to improve immersion through ambisonics and binaural development. This included the use of RIR measurement techniques and other geometric acoustic modelling techniques for representing the auralisation of a physical space. This chapter also highlighted prior research around the area of the hypothesis, discussing what had already been explored and achieved and where the research of this thesis expanded upon these previous ideas, thus demonstrating the initial novelty of the work.

Chapter 3 - Effectiveness of simple audio recreation

Chapter 3 outlined the motivation and implementation of a study titled "Real-world audio experiences can be transparently recreated in virtual reality", which explored the viability of simple methods to record the impulse response of a highly reverberant indoor environment and an open outdoor environment to authentically recreate these real-world spaces virtually. To investigate this, three audio sources were recorded in-situ and three matching audio outputs were computer generated through convolution with the RIR of both locations. Participants were asked to listen to pairs of audio recordings and determine which one was a live recording, with their responses collected anonymously through a questionnaire, along with their thoughts on the immersivity of the experience as a whole.

Potential reasons behind why the results for one specific scenario recording was so different to the rest were assessed. Real audio spaces were successfully recreated virtually to such a degree that on average, participants were more likely to favour the falsified recording in four out of the six scenarios provided. It was noted by the majority of participants that for each pair of live and falsified recordings compared, there were notable audible differences between them, but this elicited different arguments behind different participants' responses. This resulted in many respondents justifying their answers for opposing results with the same information. Potential improvements to the study were then suggested. This study marks one of the novel aspects of this thesis: tasking participants with determining which of two recordings is live and which is generated from generic acoustic data for localised sound sources.

Chapter 4 - Measuring and modelling physical spaces virtually

Chapter 4 examined an immersive way to explore the creation and exploration of auralisations for a reverberant space in an interactive virtual environment. Resonance Audio was utilized alongside CAD software to create a scale visual and aural model of Royal Holloway's Chapel to be explored in virtual reality. The study, titled "Sound produced from acoustic modelling cannot be correctly distinguished from sound produced from real-world measurements in reverberant virtual spaces", encouraged participants to explore the main body of the chapel and listen to two sound sources, a four-part choral group and a solo speaking voice. These sound sources were altered for the listener in two ways: using a pre-loaded RT60 generated from an RIR recorded at the real location, and through using Resonance Audio's reverb baking routines which use a geometric ray-traced method to pre-compute reverb for a space from geometric and material properties.

Implementation of an RT60 generated from a real-world RIR measurement into a virtual scene through Resonance Audio required a custom novel solution allowing direct access and editing of the internal properties of the reverb probes. Conversely, the other sound option was created solely using Resonance Audio's reverb calculation and the scale model of the chapel for comparison. Study participants select which of the two models they believe to be real, and which they feel is more realistic. In addition, this task was repeated with an exploration of whether a complimentary visual accompaniment to an interactive reverberant environment made identifying the real acoustic model of the space easier or harder compared to one that juxtaposed. When provided with the expected visuals, participants correctly identified the real acoustic model 45.5% of the time, compared to only 31.8% of participants when the visuals contradicted the audio.

Participants stated that they found the second scenario harder when both determining the real model and also when identifying which model they thought was more realistic for a reverberant space. These results demonstrated that sound produced from acoustic modelling could not be correctly distinguished from sound produced from real-world measurements for a reverberant space. One of the main novelties of this experiment was the participant's ability to move freely within the virtual environment while listening to various directional sounds.

Chapter 5 - The design and implementation of a real-time audio convolution system

Chapter 5 describes the design and implementation of a novel, real-time audio convolution system called The Bela Box that allows for the real-time manipulation of a live input from a performer. It grants the user the ability to alter the virtual performance space in which their voice is being simulated in. This device is portable, requiring only a power source to function. A discussion surrounding the purpose of The Bela Box and what it needed to achieve was had, highlighting the desire for use in research surrounding singing in virtual spaces. Two programming languages were explored in an effort to create the real-time convolution system, Pure Data and C++. Pd became inappropriate for the task, and C++ was utilised in tandem with Bela, a computing platform for creating low-latency audio projects, to develop the Box. Created is a compact device that convolves multiple RIRs, which can be switched between in real-time, with a live audio input. The user can also alter the ITDG of their convolution, and add dry audio if required. All of these variables can be altered while the user is performing.

Chapter 6 - Some sound cues contribute more to a person's perception of a space than others

Chapter 6 presents an experiment exploring which sound cues contribute most to a singer's perception of a space when recreating acoustics. Singers were invited to sing along with a virtual choir within a 24.1 piece speaker rig utilising a virtual reality headset and Bela Board. The results of this study provided an indication that some audio cues were indeed more important than others when trying to immerse a singer within a virtual choir, that reverberation quality was maintained in the convolution, and that the setup utilised did immerse participants in the experience.

Singers were able to alter the way their voice sounded in real-time using a novel voice altering device developed for this study, named The Bela Box, which allowed for changes in the length of the reverberant tail and the ITDG while they were singing. An assessment of a number of performance attributes, such as the ability to hear own voice and the ability to keep tempo, were posed to the singers. These results were then analysed through mean scores and standard error, demonstrating that on average, answers to all statements were overwhelmingly positive.

It was shown that participants with prior experience singing in a group choral setting in the real location used for the study were, on average, better able to pick out the correct values for the audio cues, resulting in five of the six choosing the correct reverberant tail and correct ITDG, and four of which determining both correctly. 90.5% of singers rated their enjoyment of the experience very highly, and those with prior involvement in choral singing at the original recording location stated how the set-up was “eerily accurate” and that the experience emulated the reverberation of the chapel well.

7.2 The hypothesis revisited

The level of immersion experienced by a user in virtual reality, both in terms of realism and naturalness, can be improved by implementing acoustics that are directly based on the real-world environments they are trying to represent, however, the expectations of the user also have an effect on what makes an environment sound real to them.

The hypothesis was investigated through:

1. The authentic recreation of two real-world spaces as a virtual environment that feels natural and could be acoustically explored via three distinct audio inputs: a male voice, a female voice, and an orchestral recording.

    - Results from the experiment demonstrated the concepts to be effective as in the majority of scenarios, participants struggled to correctly distinguish the live audio recordings, favouring the falsified ones for both indoor and outdoor recordings. 79.4% of individuals felt fully immersed in the chapel scenarios, and 55.9% fully immersed in the meadow scenarios.

2. The comparison of spectrograms of live acoustic measurements compared to ones falsified through convolution of Room Impulse Responses (RIRs) and Facebook 360 software.

    - Although the falsified acoustic spectrogram has a slightly less defined waveform when compared to the live recording, the recordings showed comparable signals, which was backed up by the results of the anonymous questionnaire. Participants felt as though both acoustic models could be real, with a chi-square test proving that the votes for the live and falsified recordings for all six pairs were within margin of error of each other. This supported the chapter hypothesis that real-world audio experiences could be recreated to such a degree that VR users struggled to distinguish live from falsified performances, which in turn, supports the thesis hypothesis as a whole.

3. A life-sized virtual model of Royal Holloway's chapel that could be freely explored in virtual reality, within which two auralisations could be applied in real-time, affecting audio sources in the space.

    - Participants who engaged with this study could not distinguish between auralisations produced from real-world measurements when compared to modelled ones for a reverberant indoor environment. Furthermore, when presented with a conflicting visual outdoor environment instead of the expected Chapel architecture, subjects struggled even more to determine the genuine model. The acoustic models both demonstrated the fullness of the space aurally.

4. The creation of a novel real-time voice altering device to improve immersivity in a virtual environment by adding the ability to hear reactive live audio as though the user was singing at a different location.

    - The creation of The Bela Box grants a user the ability to alter the acoustics of a virtual performance space via manipulation of the ITDG and length of reverberant tail through multiple RIRs convolved with their live audio input.
    - Singers who sang in the study were anonymously questioned about their experience, and were asked to rate specific perceptual features of their experience singing in the virtual environment. 73.9% of all participants were able to correctly identify

the correct time region of the first reflection delay, and 34.8% identified the correct reverberant tail of the impulse response. Participants commented on how the reverb was "very natural", with almost every participant rating their enjoyment of the experience highly.

5. Asking singers who had familiarity singing in a group setting to talk about the similarities with the real singing environment.

    - Participants who had prior experience singing with the same virtual choir rated the similarity to singing with them in real life as 3.67 out of 5, where 5 was the highest similarity rating. All three of these singers correctly identified both of the correct dial values.

7.2.1 Novelty of research

The main novelty of this work can be broken down into three main components:

Research probing high-end expert listener bases who are comfortable using a VR headset.

This body of work frequently engaged with individuals who were familiar with VR, most notably in the study in Chapter 3. As discussed in that chapter, it is rare that research in this field benefits from participants who are comfortable in VR [120] [57] [147]. Running this experiment with 10 VR-familiar pilot participants, followed by a further 34 VR users allowed the research throughout this study to be informed by those expert individuals. 79.4% of the participants agreed that the dynamic acoustics of the scenarios in the Chapel helped to immerse them within a scenario, stating that the "richer echoes", "directional aspect", and "sense of space" helped the most, even though the video component was at a low resolution. Comparatively, 55.9% of these participants felt immersed in the outdoor scenarios, with some participants stating how the less reverberant environment made it “harder to feel real”. The results throughout this study were positive, with participants as a whole choosing either acoustic model as real with equal likelihood. These results went on to influence later chapters, with further attention being given to investigating reverberant spaces and how realistic they would need to be when recreated virtually for a person to feel as though they were there.

Investigation of Resonance Audio as a tool for acoustic modelling of real spaces virtually.

A novel application of Resonance Audio was implemented during the research in Chapter 4. It was discovered that the RT60 values for the frequency bands of the reverb probes that determine the acoustic properties for a virtual area were stored in the file `ResonanceAudioReverbProbe.prefab`. This allowed for the creation of an editable probe, where any values for the frequency bands could be inputted and implemented by resonance audio when creating reverb properties for the space. This allowed for real-world RT60 measurements of an environment to be directly imported into the virtual equivalent. Further investigations highlighted that claims made by the developers of Resonance Audio were not repeatable in practice. This includes a misleading demo where users can explore the use of reverb probes over three acoustically different environments, one of which is a cathedral. The pre-baked RT60 for this reverberant environment is very dynamic, more so than the equivalently sized chapel created for the experiment in Chapter 4. However, on re-baking the reverb probe, the new RT60 frequency band values drop significantly. Similarly, the frequency band values that resonance audio created for the model of the Chapel in this study were also lower than anticipated by the real-world measurements. Future work should be undertaken to explore these discrepancies in what is promised, and what is produced by Resonance Audio.

The Bela Box.

Finally, The Bela Box is a portable device which can convolve real-time audio inputs from a performer with one of up to five IRs. The ITDG of these IRs can also be decreased and extended, allowing the user to alter the way their voice sounds in real time. This was achieved using a Bela board, an ultra-low latency device designed for audio programming. C++ was used to create both a physical and graphical user interface, with the physical interface having been integrated with a VR headset to create an immersive singing experience. Singers with familiarity singing in a group setting in the chapel described the reverb as "very natural", and that "singing with the virtual choir was quite realistic to singing with the live choir". This device is not only novel, but it is key in helping to prove what the initial hypothesis set out to achieve by fully integrating acoustics from a real-world environment to enhance the naturalness of real-world singing virtually.

7.3 Further work

7.3.1 Further development for real-time singing experiences

Personalisation for virtual singing environment

There are multiple minor alterations to be executed to improve overall immersion when singing with a virtual choir. Firstly, personalized Head Related Transfer Functions for the off-ear speakers for the scenario where a 24.1-piece speaker rig is unavailable. As shown through research into Vive Cinema and the work done by the SADIE project, the ability to implement various HRTFs is attainable for similar software and improves personal engagement when immersing somebody fully in a virtual environment [55] [159] [214]. Therefore, it would be advantageous to tailor the personal singing experience for each participant if, for example, the speaker rig was not available, and the experience needed to be portable.

Secondly, regarding the hypothesis investigated in Chapter 6, a general assumption based on previous research was made when thinking about a person's ability to discern latency. The work suggested that on average, a delay of 3 to 10ms can be present without being noticed by an individual. Running a pre-experiment prior to the study to determine a participant's sensitivity to latency would have allowed for discussion around the importance of latency of the first reflection to be tailored to personal ability, providing a more accurate overview of the results.

Further comparison to real-world singing locations

Briefly touched upon in Chapter 6, some of the statements posed to participants allowed for reflection on their own confidence in their ability instead of how the virtual singing environment felt to them compared to real life. Therefore, when further exploring the groundwork cemented in this study, it would be valuable to have singers first perform in the real physical environment as well as the virtual one, and have questions explore the comparison of the two experiences. This could allow for the measurement of other factors, such as loudness or intonation comparatively in each environment.

Location of listening position when singing

The ability to alter the ITDG using the Bela Box highlights some interesting potential other uses for a singer using the device. For example, future work could seek to include the opportunity for a singer to change the position from which they are hearing themselves as a member of an audience or from the conductor's position. This would allow for a performer to monitor their own performance in the virtualised version of a location in real time, which could be a great tool for singing training and rehearsals in unfamiliar acoustic environments.

7.3.2 Improving the fully explorable virtual environments

Visuals for a virtual recreation of a real environment

As mentioned in Chapter 4, LiDAR scanning technologies can be used to relatively quickly create accurate 3D physical models of spaces that can be combined with 2D imagery to construct photorealistic models. This would both speed up the creation of physical environments for VR as well as improve immersivity with photorealism.

Using the Bela board for virtual reality tracking

The Bela board can be employed alongside a head tracking sensor to create immersive audio scenes that are explored through head movements. Therefore, it would not be unreasonable to assume that the board could also be used in tandem with a VR headset, like in Chapter 6, but where the acoustics are truer to the space, and change based on head orientation and direction. The impulse response recorded was for a single position and location, however through Bela, there could be potential for a more acoustically interactive virtual environment that involves a dynamic application of an auralisation of a space.

This area of research could, for example, allow for a collaboration between what was achieved in the study surrounding the physical exploration of the Chapel and the study where participants sang along with a choir, adding real-time audio inputs to an explorable virtual environment. There have not been any truly open-world projects wherein a person could perform freely anywhere in a virtual space and have the convolution update in real-time based on where in the environment the performer is.

7.4 Applications of research

7.4.1 Real-time convolution

A novel real-time convolver was developed as part of this thesis, which will inform other applications that utilise real-time auralisations for research. Use of the Bela board helped to cement its usefulness within ultra-low latency work, which will hopefully highlight its capabilities for research purposes as well as for immersive art. An exploration of how correct the ITDG values need to be for a person to feel as though the acoustics for a space are the same as their real-world alternative was achieved, which demonstrates a window wherein unhelpful latency could be absorbed within a process. This project used five impulse responses on a single board; other studies could explore more reverberant spaces up to five times the length of reverberant tail that was implemented in this work.

7.4.2 Rehearsal and performance

If the Bela board in conjunction with either off-ear speakers and/or a surround speaker rig simulated the acoustics of a performance space well, then there is potential for use as a tool for rehearsal and performing in general. This is a multifaceted application. Firstly, if for whatever reason a person could not get to a group rehearsal for an extended period of time, they could still gain the experience of singing within a group at a desired location. Secondly, these systems could be utilised as a rehearsal tool where the final performance location is inaccessible in some way, either due to travel requirements, scheduling, or financial limitations. A performer could load a virtual version of a performance environment, both visually and aurally, and rehearse as though they are actually there, merging the ability to freely walk around the space with a real-time, adaptive convolution of their voice.

7.4.3 Psychoacoustics

Perception of sound and perception of immersion for a person was explored in all three major experiments conducted, informing the field of psychoacoustics through insight into what a person requires to perceive themselves as immersed in a virtual environment. Chapter 6 especially helped to identify key acoustic cues which influence a singer's ability to perform naturally in a modelled acoustic space. There is room for further exploration here, for example, including further adaptable acoustic parameters to explore other facets that affect immersivity for the virtualisation of performance spaces.

7.4.4 Virtual reality and immersive audio

This research highlighted and utilised free, research-based software throughout the entire project. From Vive Cinema and its ability to play 360° videos with ambisonic audio and variable HRTFs [159], COMPASS and its many tools such as the ambisonic upmixer [210], and Angelo Farina's Aurora plug-in suite [182], this research informs future researchers of many excellent tools present to aid with their own research. Furthermore, novel methods of auralisation and exploration of immersive audio were also undertaken in this research, such as the creation of the Bela Box with its interactive real-time, low-latency voice altering capabilities.

7.4.5 Architectural design past and present

The ideas explored in this research could be used for exploration into past and future performance spaces. Locations that are in partial or full ruin have been virtually rebuilt in other work [174,177,188]. However in all these cases, the ability to move freely and perform within the virtual space is still not implemented. Doing this would allow performers and researchers to hear how these ancient spaces would have sounded when in their full glory in all areas of the space. Furthermore, it could highlight potential design choices to improve spatial acoustics that we may have neglected over the years. Finally, some concepts suggested could be applied to current performance space design, allowing for architects to test out the acoustics of a potential performance environment prior to building.

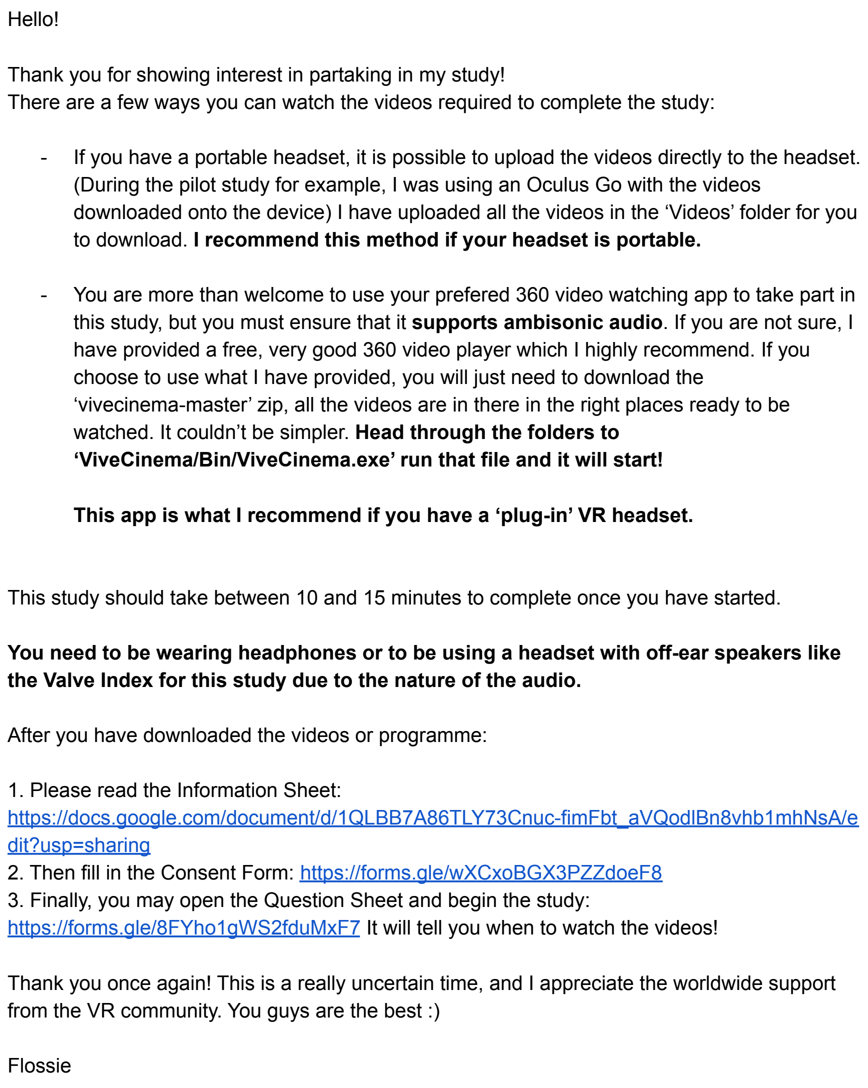
7.5 Final comments

The research presented in this thesis has investigated the realities of virtual voice production in performance spaces through the creation of novel devices and questions whilst facilitating future research areas. The results gathered in this thesis highlight which aspects of acoustics within a virtual environment need to be identical to that of the real space, and which are suitable when similar enough. The most important of these occurring when a user can provide their own stimulus in real time into an aural virtual environment, such as through singing. When presented with sounds to listen to, even with the ability to move freely about a virtual space, the inconsistencies between models and real life were shown not to make a significant difference in a person's immersion.

Though the techniques and research presented in this work do not fully outline every facet of what is required to naturally recreate the acoustics of spaces virtually, they do offer progress and some certainties. It is hoped that the research presented in this thesis can offer a basis from which others may investigate aural cues and qualities for immersive virtual environments in VR. Chapters 3 to 6 highlight both well known and relatively unknown devices and programs that were key in the implementation of much of this work, which hopefully can provide a thorough overview of potential direction, both for hardware and software, for future research endeavours. Let us create a greater immersion when exploring virtual spaces, whether it is for entertainment as a way to lose oneself in a wondrous world, or for professionals learning and working on real tasks virtually. As the founder of Oculus once said, "why shouldn't we be able to teleport wherever we want? [37]".

Appendix A

Instructions and Questionnaire for Study 1 (Chapter 3)



Hello!

Thank you for showing interest in partaking in my study!
There are a few ways you can watch the videos required to complete the study:

- If you have a portable headset, it is possible to upload the videos directly to the headset. (During the pilot study for example, I was using an Oculus Go with the videos downloaded onto the device) I have uploaded all the videos in the 'Videos' folder for you to download. **I recommend this method if your headset is portable.**

- You are more than welcome to use your prefered 360 video watching app to take part in this study, but you must ensure that it **supports ambisonic audio**. If you are not sure, I have provided a free, very good 360 video player which I highly recommend. If you choose to use what I have provided, you will just need to download the 'vivecinema-master' zip, all the videos are in there in the right places ready to be watched. It couldn't be simpler. **Head through the folders to 'ViveCinema/Bin/ViveCinema.exe' run that file and it will start!**

  This app is what I recommend if you have a 'plug-in' VR headset.

This study should take between 10 and 15 minutes to complete once you have started.

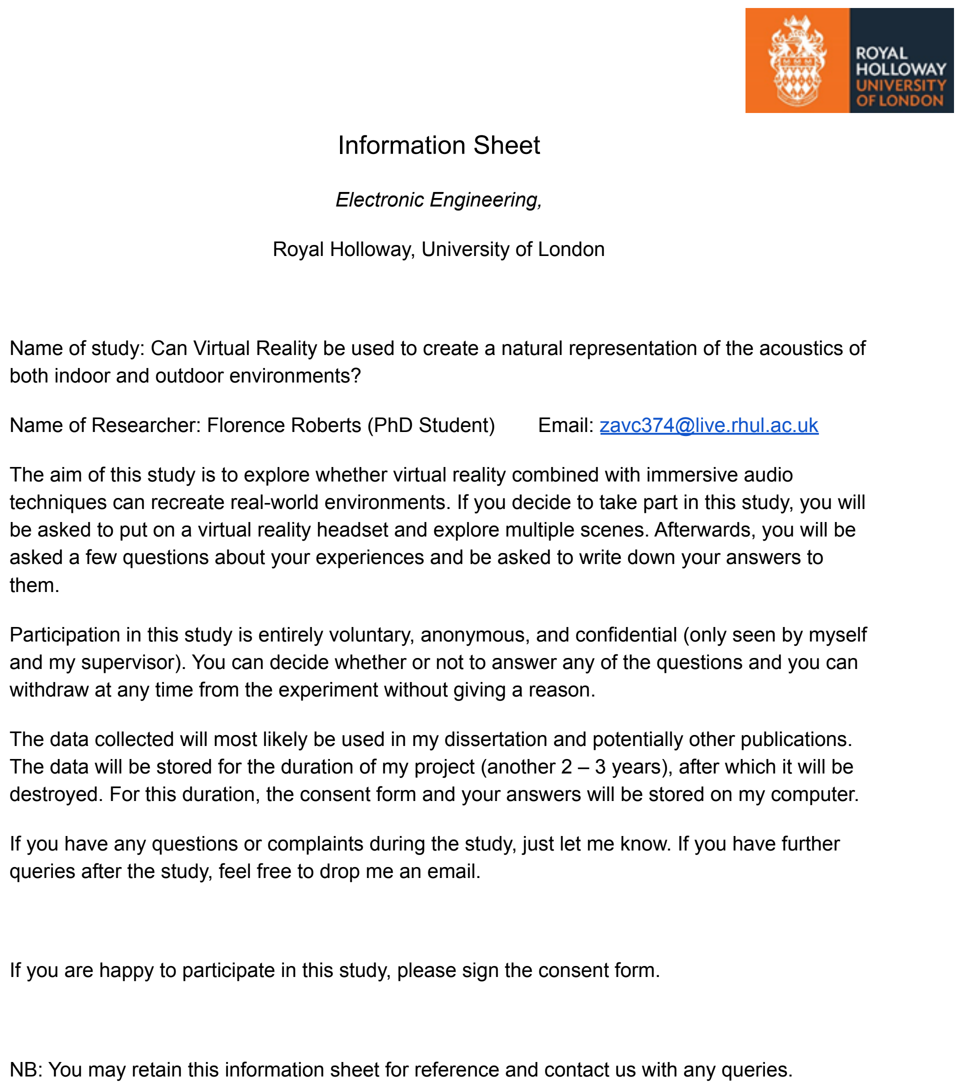
You need to be wearing headphones or to be using a headset with off-ear speakers like the Valve Index for this study due to the nature of the audio.

After you have downloaded the videos or programme:

1. Please read the Information Sheet:
https://docs.google.com/document/d/1QLBB7A86TLY73Cnuc-fimFbt_aVQodlBn8vhb1mhNsA/edit?usp=sharing
2. Then fill in the Consent Form: https://forms.gle/wXCxoBGX3PZZdoeF8
3. Finally, you may open the Question Sheet and begin the study:
https://forms.gle/8FYho1gWS2fduMxF7 It will tell you when to watch the videos!

Thank you once again! This is a really uncertain time, and I appreciate the worldwide support from the VR community. You guys are the best :)

Flossie



Information Sheet

Electronic Engineering,

Royal Holloway, University of London

Name of study: Can Virtual Reality be used to create a natural representation of the acoustics of both indoor and outdoor environments?

Name of Researcher: Florence Roberts (PhD Student)   Email: zavc374@live.rhul.ac.uk

The aim of this study is to explore whether virtual reality combined with immersive audio techniques can recreate real-world environments. If you decide to take part in this study, you will be asked to put on a virtual reality headset and explore multiple scenes. Afterwards, you will be asked a few questions about your experiences and be asked to write down your answers to them.

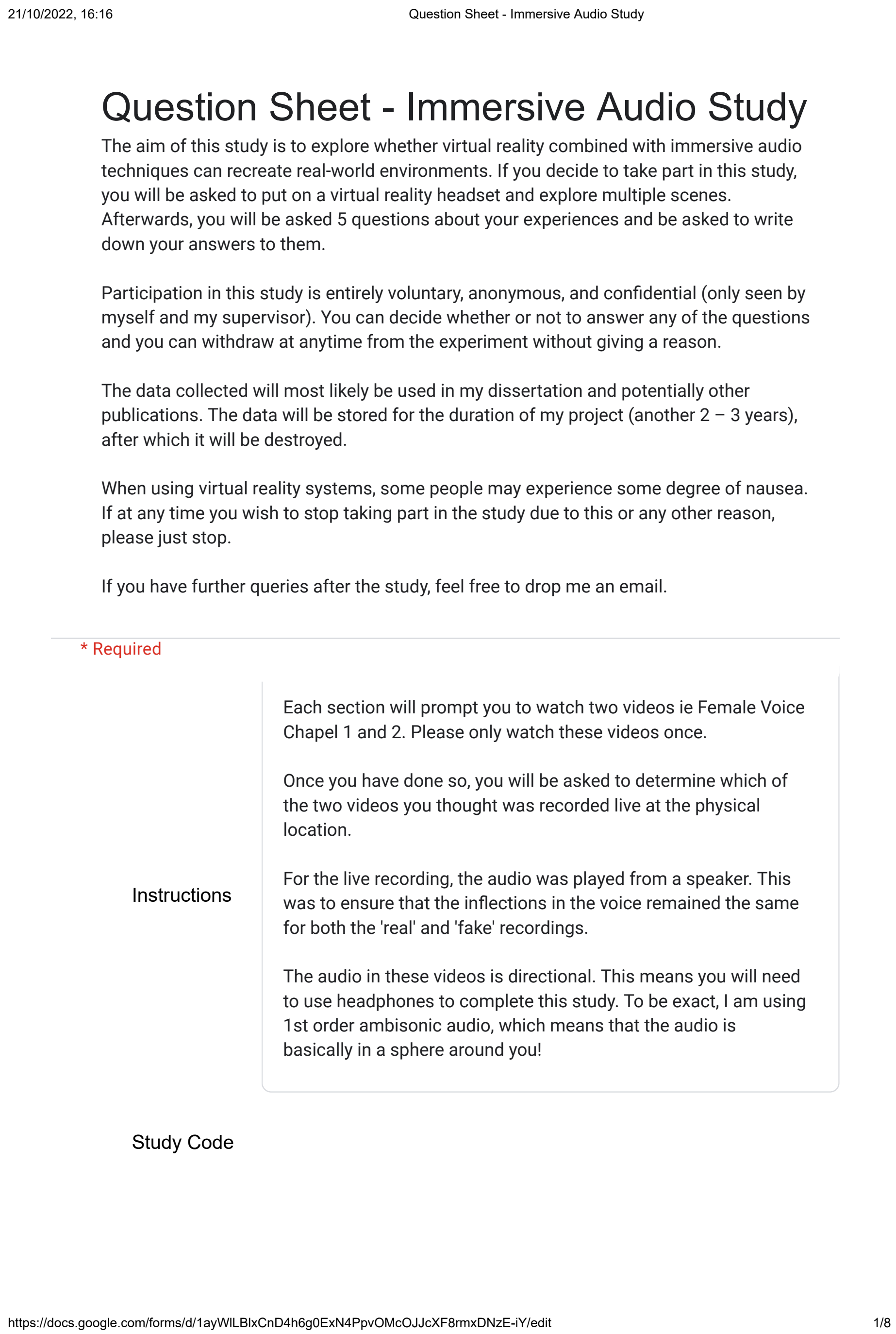
Participation in this study is entirely voluntary, anonymous, and confidential (only seen by myself and my supervisor). You can decide whether or not to answer any of the questions and you can withdraw at any time from the experiment without giving a reason.

The data collected will most likely be used in my dissertation and potentially other publications. The data will be stored for the duration of my project (another 2 – 3 years), after which it will be destroyed. For this duration, the consent form and your answers will be stored on my computer.

If you have any questions or complaints during the study, just let me know. If you have further queries after the study, feel free to drop me an email.

If you are happy to participate in this study, please sign the consent form.

NB: You may retain this information sheet for reference and contact us with any queries.



21/10/2022, 16:16 Question Sheet - Immersive Audio Study

Question Sheet - Immersive Audio Study

The aim of this study is to explore whether virtual reality combined with immersive audio techniques can recreate real-world environments. If you decide to take part in this study, you will be asked to put on a virtual reality headset and explore multiple scenes. Afterwards, you will be asked 5 questions about your experiences and be asked to write down your answers to them.

Participation in this study is entirely voluntary, anonymous, and confidential (only seen by myself and my supervisor). You can decide whether or not to answer any of the questions and you can withdraw at anytime from the experiment without giving a reason.

The data collected will most likely be used in my dissertation and potentially other publications. The data will be stored for the duration of my project (another 2 – 3 years), after which it will be destroyed.

When using virtual reality systems, some people may experience some degree of nausea. If at any time you wish to stop taking part in the study due to this or any other reason, please just stop.

If you have further queries after the study, feel free to drop me an email.

* Required

Instructions

Each section will prompt you to watch two videos ie Female Voice Chapel 1 and 2. Please only watch these videos once.

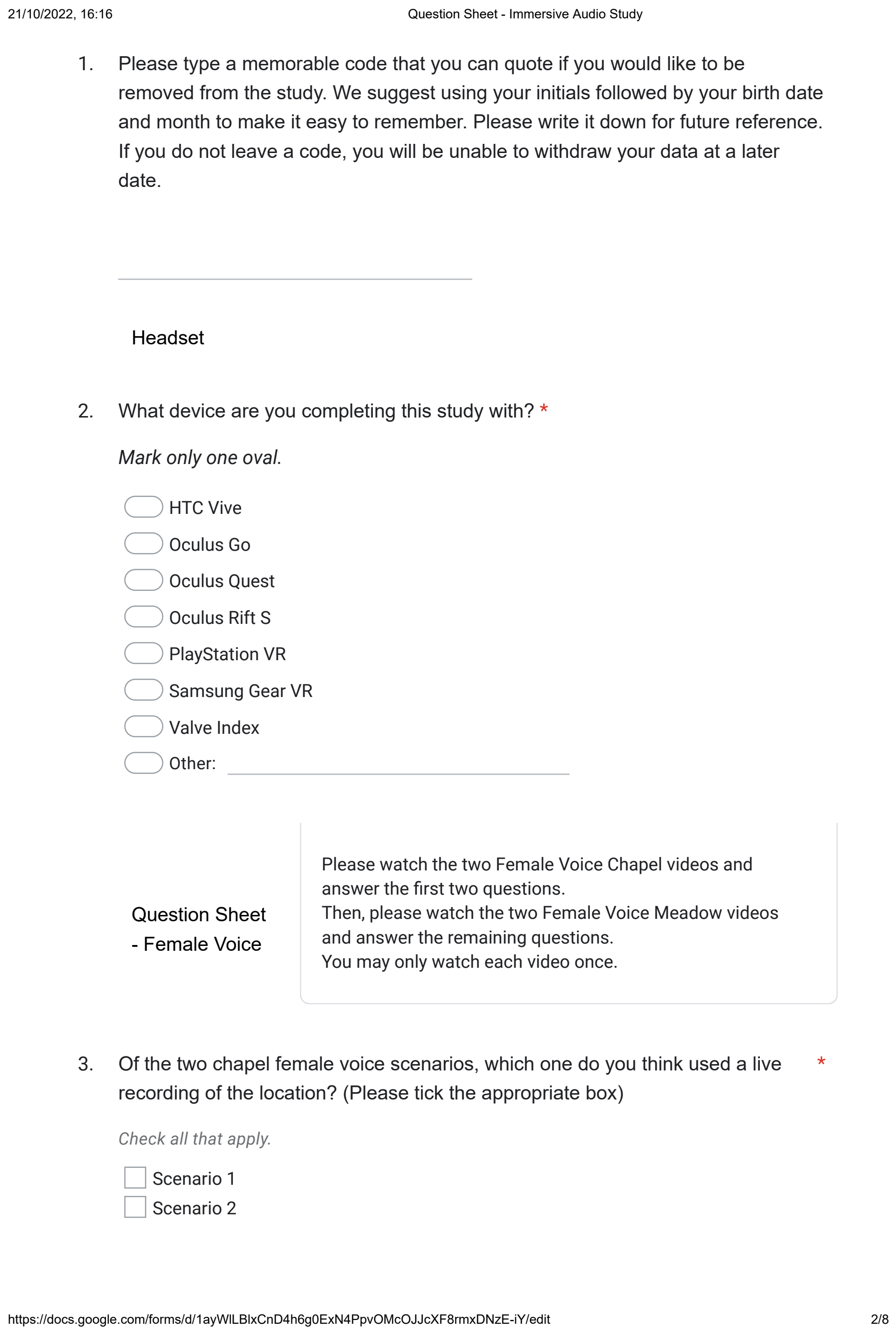
Once you have done so, you will be asked to determine which of the two videos you thought was recorded live at the physical location.

For the live recording, the audio was played from a speaker. This was to ensure that the inflections in the voice remained the same for both the 'real' and 'fake' recordings.

The audio in these videos is directional. This means you will need to use headphones to complete this study. To be exact, I am using 1st order ambisonic audio, which means that the audio is basically in a sphere around you!

Study Code

https://docs.google.com/forms/d/1ayWlLBlxCnD4h6g0ExN4PpvOMcOJJcXF8rmxDNzE-iY/edit 1/8



1. Please type a memorable code that you can quote if you would like to be removed from the study. We suggest using your initials followed by your birth date and month to make it easy to remember. Please write it down for future reference. If you do not leave a code, you will be unable to withdraw your data at a later date.

____

Headset

2. What device are you completing this study with? *

Mark only one oval.

○ HTC Vive
○ Oculus Go
○ Oculus Quest
○ Oculus Rift S
○ PlayStation VR
○ Samsung Gear VR
○ Valve Index
○ Other: ______________________________

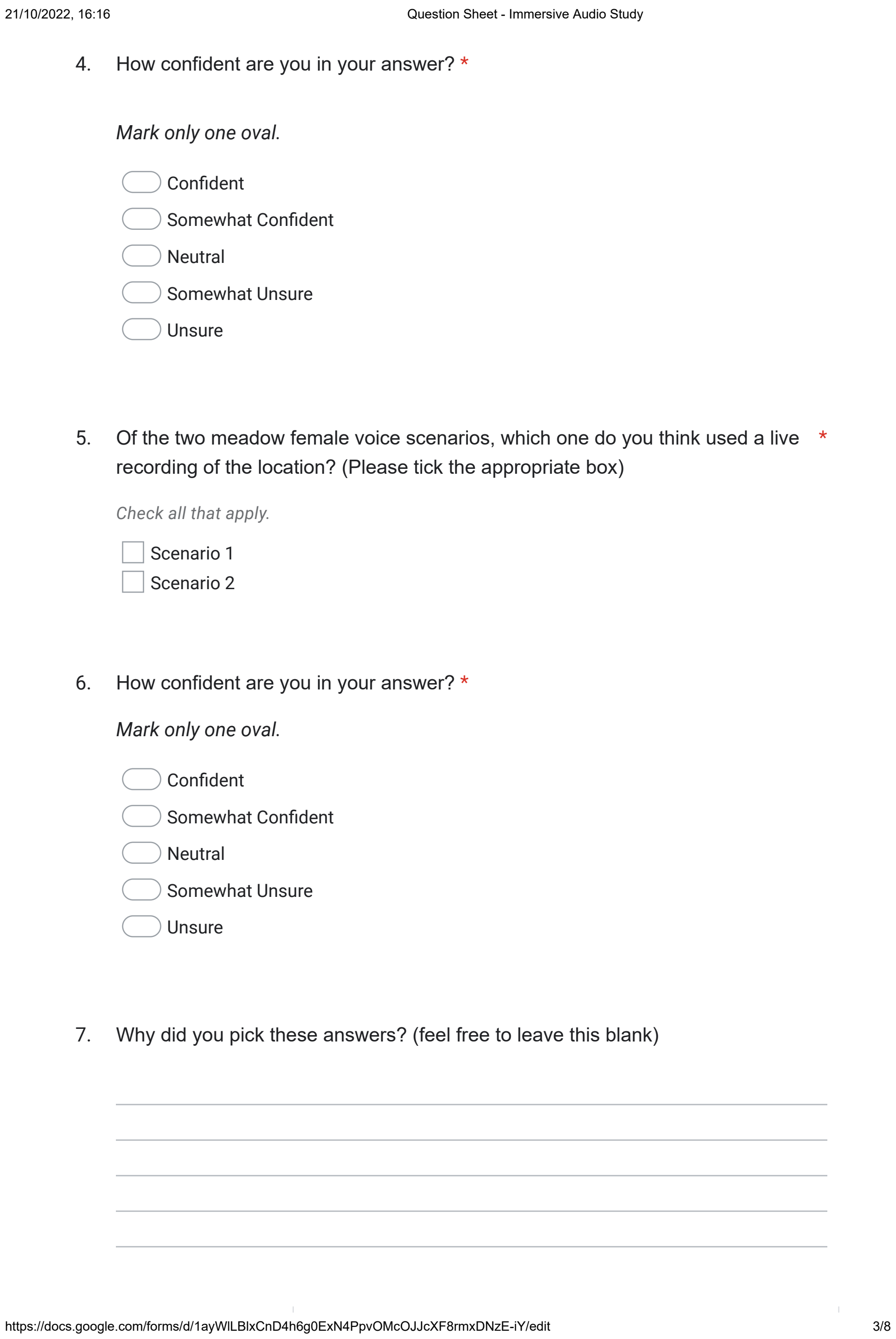
Question Sheet - Female Voice

Please watch the two Female Voice Chapel videos and answer the first two questions.
Then, please watch the two Female Voice Meadow videos and answer the remaining questions.
You may only watch each video once.

3. Of the two chapel female voice scenarios, which one do you think used a live recording of the location? (Please tick the appropriate box) *

Check all that apply.

□ Scenario 1
□ Scenario 2



21/10/2022, 16:16 Question Sheet - Immersive Audio Study

4. How confident are you in your answer? *

   Mark only one oval.

   ○ Confident
   ○ Somewhat Confident
   ○ Neutral
   ○ Somewhat Unsure
   ○ Unsure

5. Of the two meadow female voice scenarios, which one do you think used a live recording of the location? (Please tick the appropriate box) *

   Check all that apply.

   □ Scenario 1
   □ Scenario 2

6. How confident are you in your answer? *

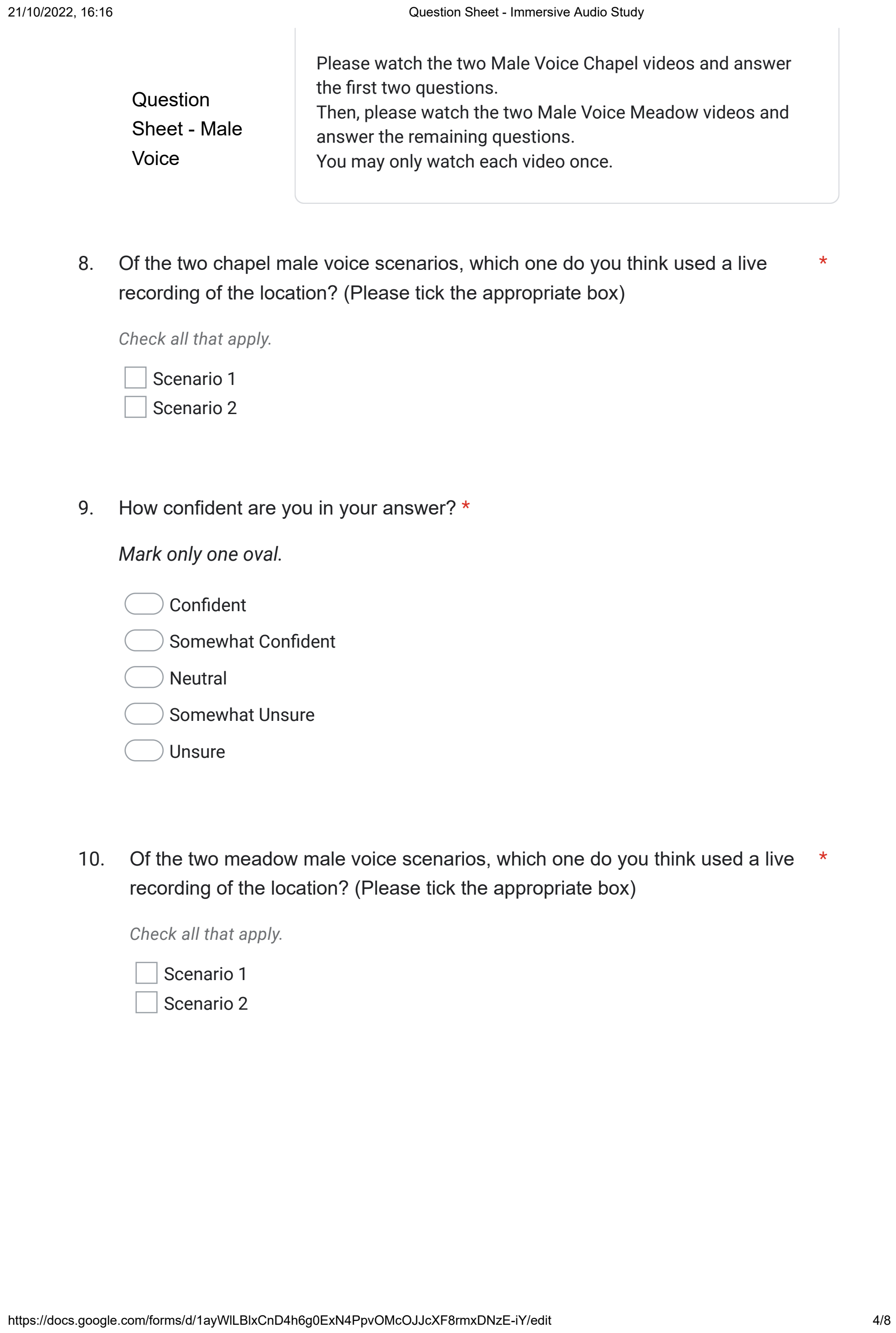
   Mark only one oval.

   ○ Confident
   ○ Somewhat Confident
   ○ Neutral
   ○ Somewhat Unsure
   ○ Unsure

7. Why did you pick these answers? (feel free to leave this blank)

   ____
   ____
   ____
   ____
   ____

https://docs.google.com/forms/d/1ayWlLBlxCnD4h6g0ExN4PpvOMcOJJcXF8rmxDNzE-iY/edit 3/8



Question Sheet - Male Voice

Please watch the two Male Voice Chapel videos and answer the first two questions.
Then, please watch the two Male Voice Meadow videos and answer the remaining questions.
You may only watch each video once.

8. Of the two chapel male voice scenarios, which one do you think used a live recording of the location? (Please tick the appropriate box) *

Check all that apply.

□ Scenario 1
□ Scenario 2

9. How confident are you in your answer? *

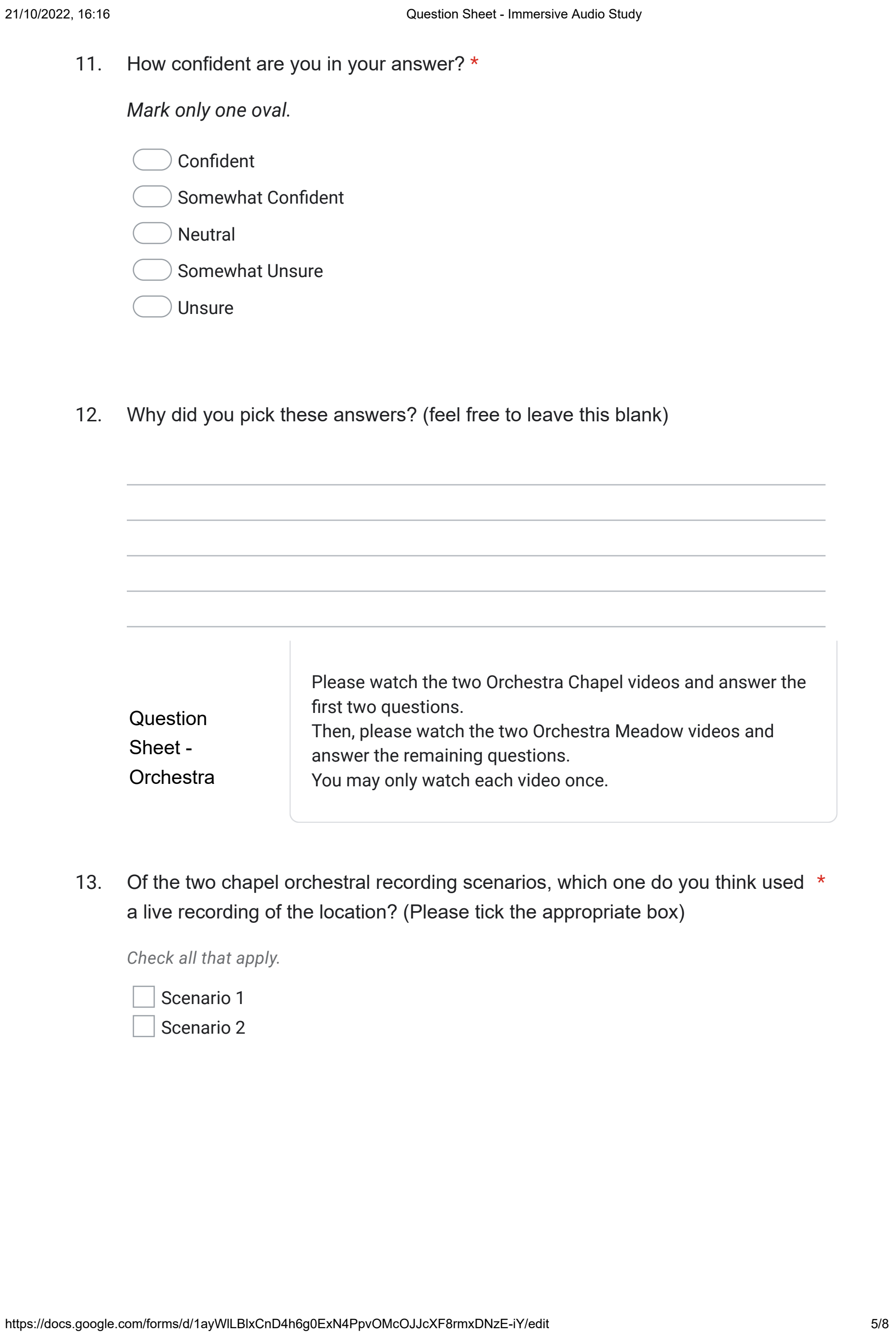
Mark only one oval.

○ Confident
○ Somewhat Confident
○ Neutral
○ Somewhat Unsure
○ Unsure

10. Of the two meadow male voice scenarios, which one do you think used a live recording of the location? (Please tick the appropriate box) *

Check all that apply.

□ Scenario 1
□ Scenario 2



21/10/2022, 16:16 Question Sheet - Immersive Audio Study

11. How confident are you in your answer? *

Mark only one oval.

○ Confident
○ Somewhat Confident
○ Neutral
○ Somewhat Unsure
○ Unsure

12. Why did you pick these answers? (feel free to leave this blank)

Question Sheet - Orchestra

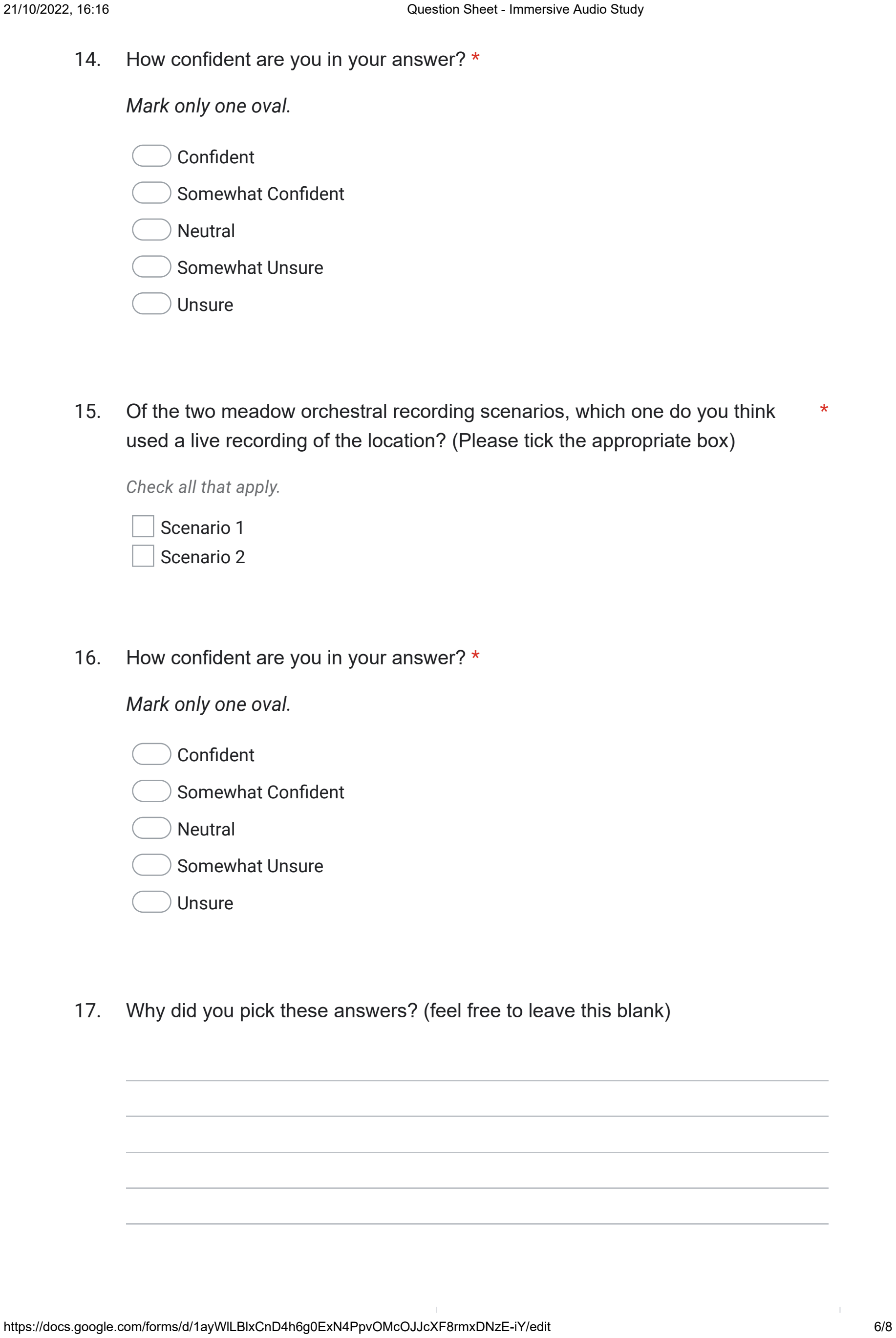
Please watch the two Orchestra Chapel videos and answer the first two questions.
Then, please watch the two Orchestra Meadow videos and answer the remaining questions.
You may only watch each video once.

13. Of the two chapel orchestral recording scenarios, which one do you think used a live recording of the location? (Please tick the appropriate box) *

Check all that apply.

□ Scenario 1
□ Scenario 2

https://docs.google.com/forms/d/1ayWlLBlxCnD4h6g0ExN4PpvOMcOJJcXF8rmxDNzE-iY/edit 5/8



14. How confident are you in your answer? *

Mark only one oval.

○ Confident
○ Somewhat Confident
○ Neutral
○ Somewhat Unsure
○ Unsure

15. Of the two meadow orchestral recording scenarios, which one do you think used a live recording of the location? (Please tick the appropriate box) *

Check all that apply.

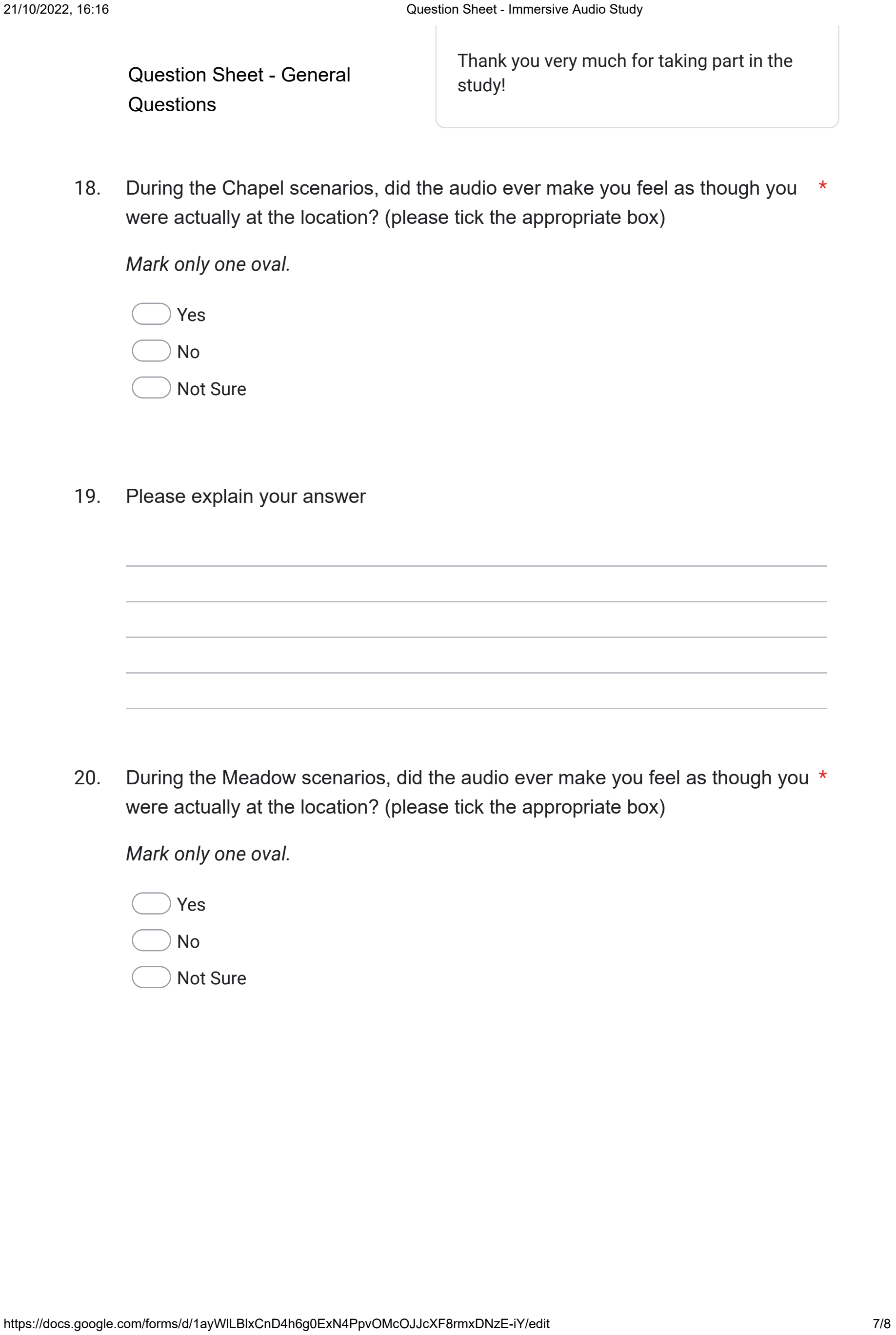
□ Scenario 1
□ Scenario 2

16. How confident are you in your answer? *

Mark only one oval.

○ Confident
○ Somewhat Confident
○ Neutral
○ Somewhat Unsure
○ Unsure

17. Why did you pick these answers? (feel free to leave this blank)



21/10/2022, 16:16 Question Sheet - Immersive Audio Study

Question Sheet - General Questions

Thank you very much for taking part in the study!

18. During the Chapel scenarios, did the audio ever make you feel as though you were actually at the location? (please tick the appropriate box) *

Mark only one oval.

○ Yes
○ No
○ Not Sure

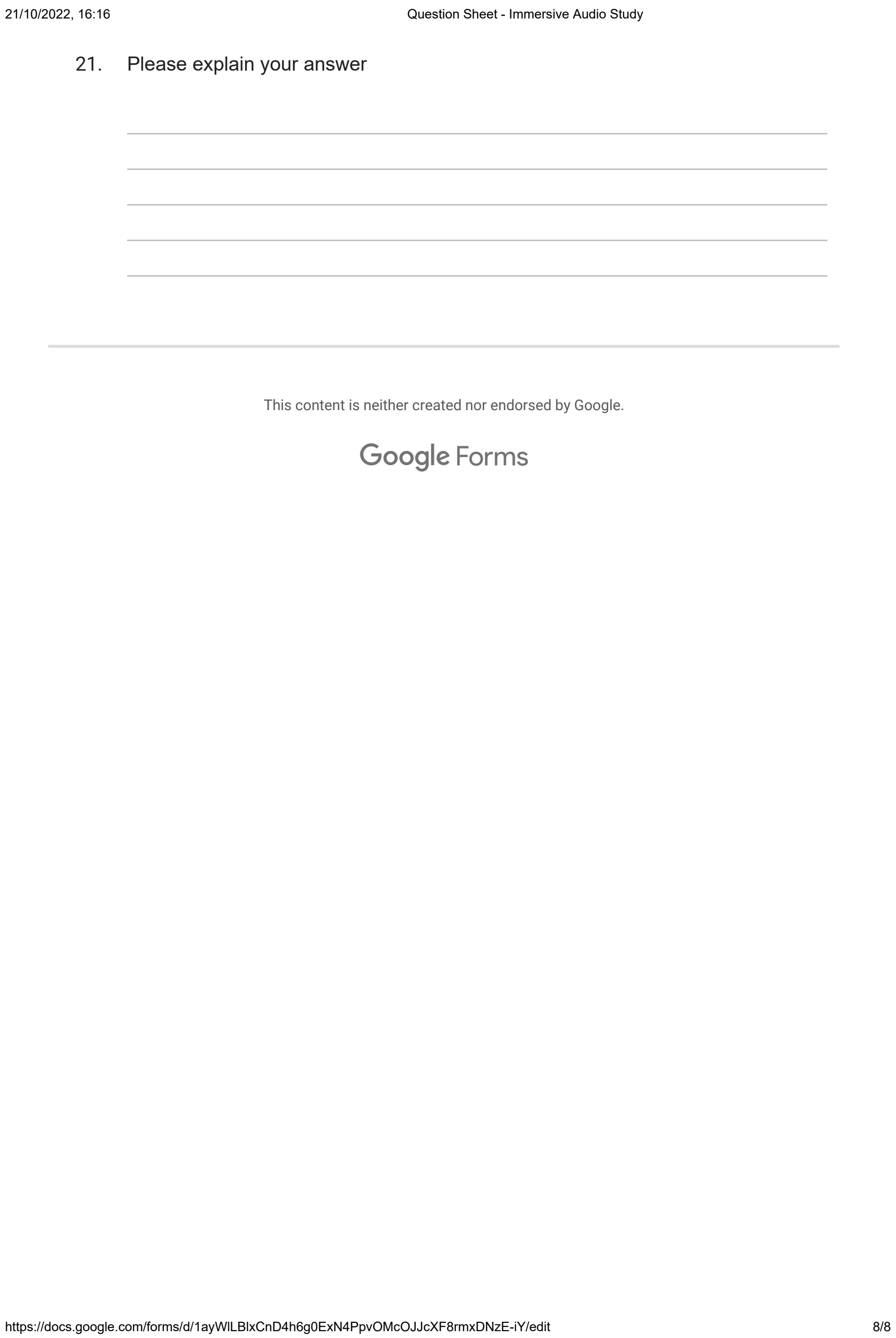
19. Please explain your answer

20. During the Meadow scenarios, did the audio ever make you feel as though you were actually at the location? (please tick the appropriate box) *

Mark only one oval.

○ Yes
○ No
○ Not Sure

https://docs.google.com/forms/d/1ayWlLBlxCnD4h6g0ExN4PpvOMcOJJcXF8rmxDNzE-iY/edit 7/8



21/10/2022, 16:16 Question Sheet - Immersive Audio Study

21. Please explain your answer

This content is neither created nor endorsed by Google.

https://docs.google.com/forms/d/1ayWlLBlxCnD4h6g0ExN4PpvOMcOJJcXF8rmxDNzE-iY/edit 8/8

Appendix B

Instructions and Questionnaire for Study 2 (Chapter 4)

Hello!

Thank you for agreeing to take part in this project!
The goal of this study is to compare the perceived validity of sound fabricated using acoustic modelling techniques and sound produced using real world measurements, in varying virtual spaces.

In this study, you will be presented with two scenarios: A and B. In each of these, you will be able to change between 2 acoustic models for a reverberant space (Royal Holloway's Chapel) and play multiple sound sources in that space.

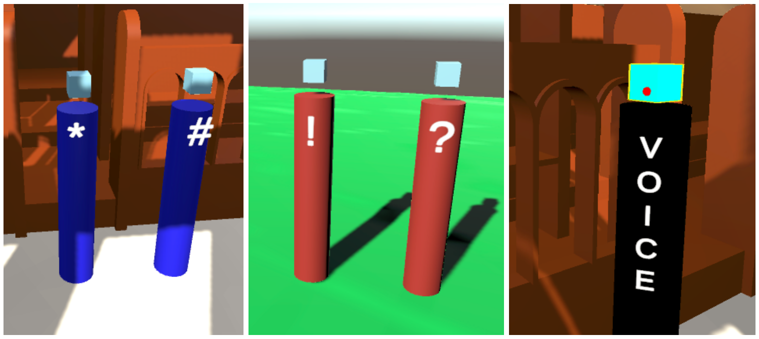
These scenarios will differ visually, but for both cases you will be asked the same thing: which acoustic model you thought was created from real world data, and which of the two you thought sounded the most realistic for the reverberant space, whatever your visual surroundings look like. When you first load in, you will appear in a Hub World that lets you learn about the controls for the study.

Instructions:

1. Please read the information sheet: https://docs.google.com/document/d/1QH4qFZxe9dA92d65oPXo9T1aqkmNhl5GOXZJkFFGJn4/edit?usp=sharing
2. Please fill in the consent form: https://forms.gle/8S9rnr8DuAkrYgMZA
3. You will need to have Steam VR downloaded and connected to your headset
4. If you are doing this remotely, you need to download everything from the 'ToDownload' zip file in the drive link you were sent. The project is run through 'Study 2.exe'.

   You also need to be wearing headphones or to be using a headset with off-ear speakers like the Valve Index for this study due to the nature of the audio.

5. When you enter the scenarios after the hub world, before doing anything else, pick one of the two acoustic models. This is done by pushing down on a button with your virtual hand, and letting go after it turns blue. Below are what the buttons for scenarios A and B look like, as well as an example of the button being pressed for reference:

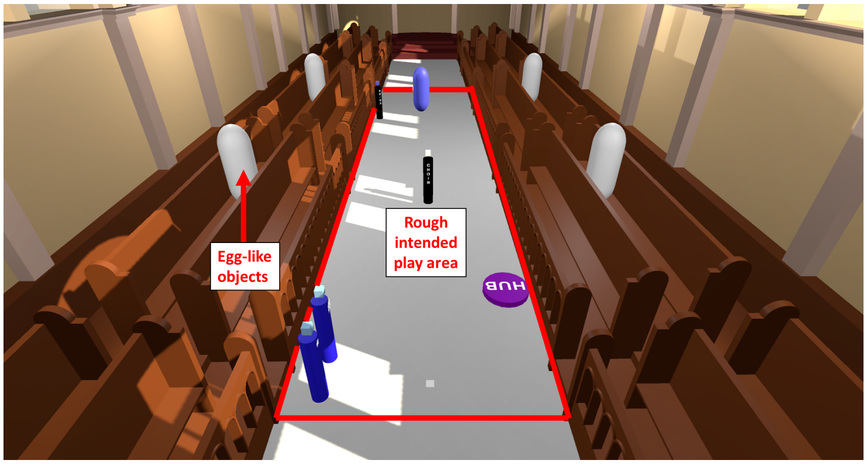


a)Scenario A's buttons **b)Scenario B's buttons** **c)Pressed button reference**

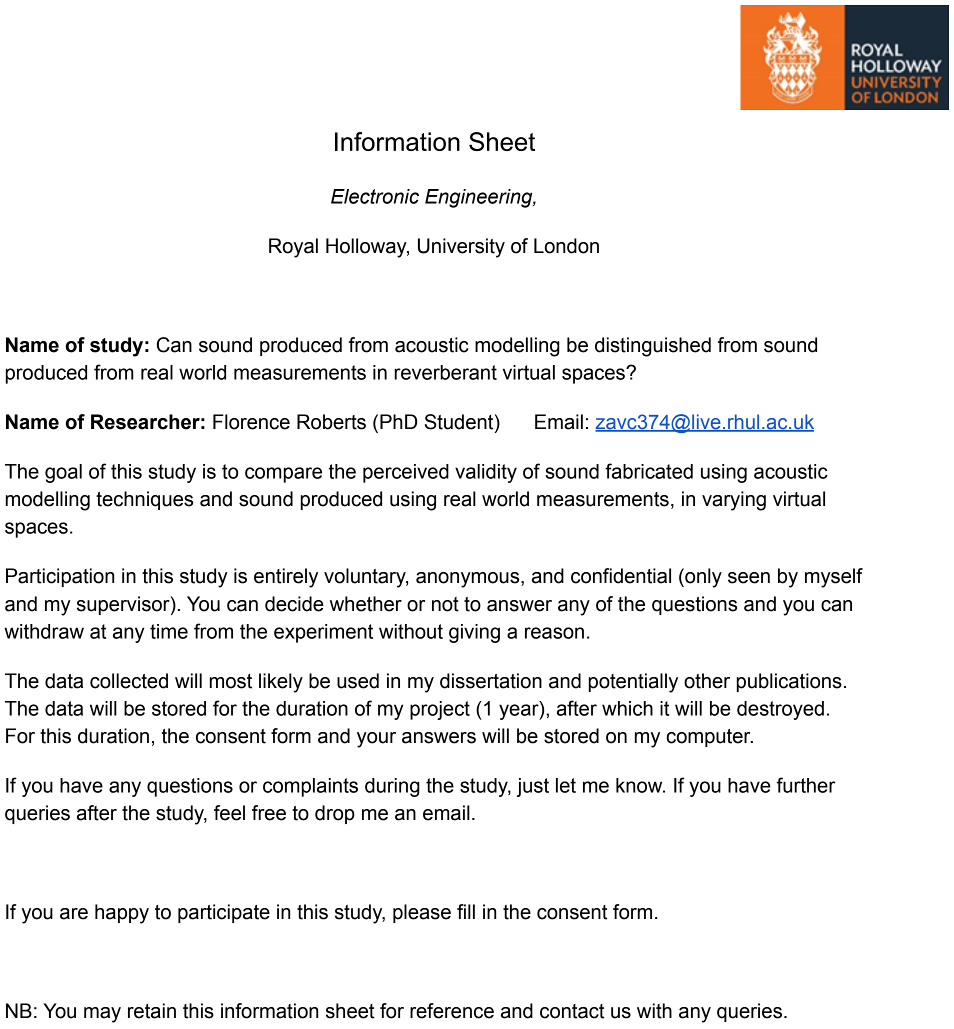
6. There are 2 sound sources for each space - a choir and a voice. These buttons are on black pillars (see fig. c). Listen to both for the 2 acoustic models as many times as you like, just make sure to interact with each combination at least once. You can change both the audio sources and the acoustic models whenever you like, including when sounds are already being played.
7. The egg-shaped objects represent the people singing and speaking, and they produce their sound like a person - they are louder on one side than the other, for example. Feel free to move around while the sounds are playing!
8. Please complete scenario A and then scenario B, stopping to answer the questionnaire about A before moving on. Do not do both at the end. Questionnaire: https://forms.gle/SAjyr2wLmfipmAzT6

Bugs to watch out for:

- The intention is that you move around in the aisle only, however while running this study, people have managed to break out of bounds. If at any point you end up in the pews, or in the case of the second scenario, near the white egg shaped objects (you'll see what I mean) you have broken out of bounds. The easiest way to see where these boundaries are supposed to restrict you is to hold the teleport button out in front or to the side of you until it lies flat against an invisible wall. Please head back in bounds as quickly as you can if this happens.
- In the Hub World on instruction 2, sometimes you need to press the start, then stop, then the start button again to make any noise appear. I don't know why, but please make sure you hear the noise before entering the scenarios so you can adjust your volume.
- Rarely, the audio crackles and slows down a bit, which can impair your experience. If this happens, play the voice button file once all the way to the end without interruption, then try the other sound file again. Sometimes you'll just need to restart the program. It is sometimes triggered by very rapid teleportation (as a heads up).



That’s it! When you are ready to start, run ‘Study 2.exe’. Thank you very much!



Information Sheet

Electronic Engineering,

Royal Holloway, University of London

Name of study: Can sound produced from acoustic modelling be distinguished from sound produced from real world measurements in reverberant virtual spaces?

Name of Researcher: Florence Roberts (PhD Student) Email: zavc374@live.rhul.ac.uk

The goal of this study is to compare the perceived validity of sound fabricated using acoustic modelling techniques and sound produced using real world measurements, in varying virtual spaces.

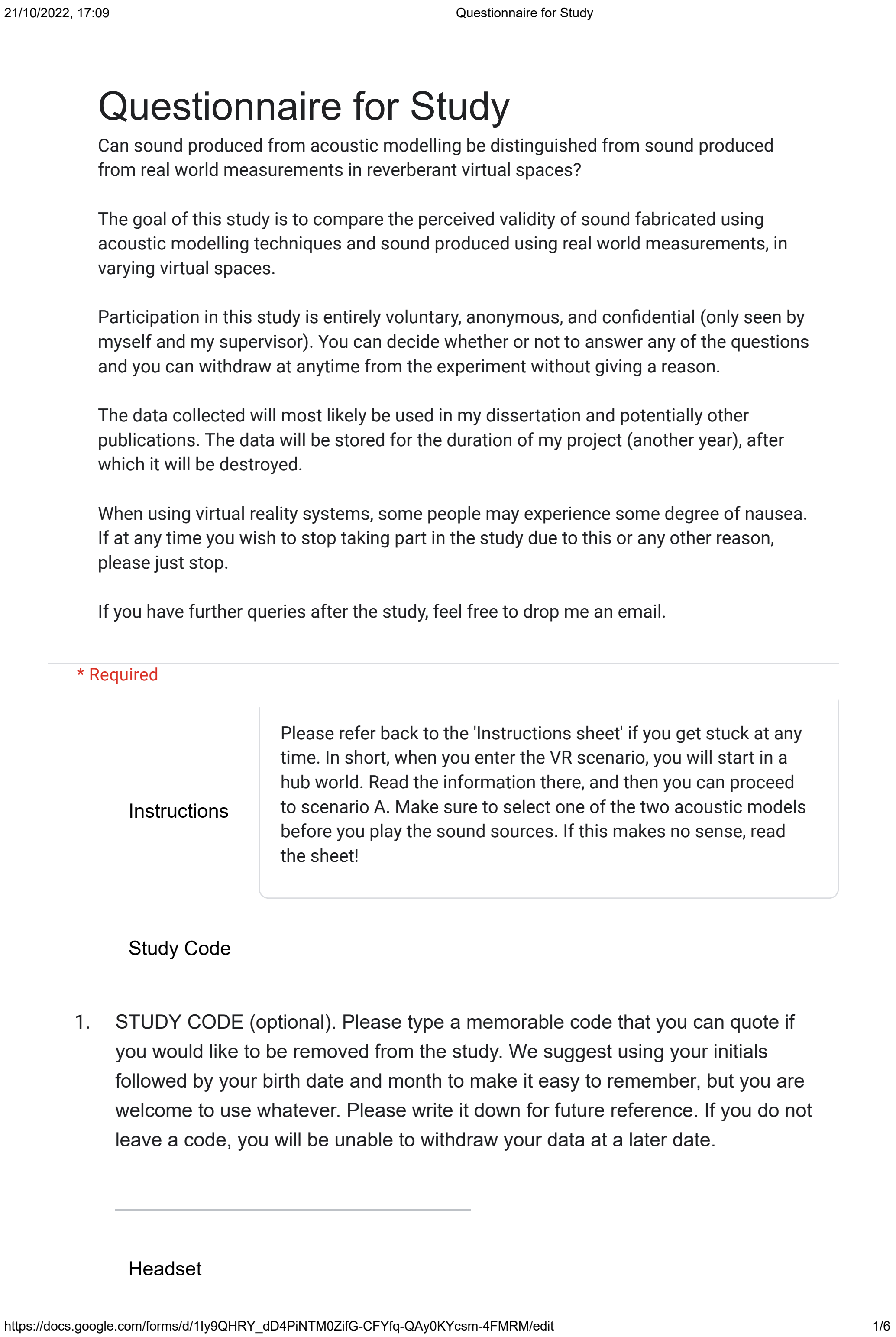
Participation in this study is entirely voluntary, anonymous, and confidential (only seen by myself and my supervisor). You can decide whether or not to answer any of the questions and you can withdraw at any time from the experiment without giving a reason.

The data collected will most likely be used in my dissertation and potentially other publications. The data will be stored for the duration of my project (1 year), after which it will be destroyed. For this duration, the consent form and your answers will be stored on my computer.

If you have any questions or complaints during the study, just let me know. If you have further queries after the study, feel free to drop me an email.

If you are happy to participate in this study, please fill in the consent form.

NB: You may retain this information sheet for reference and contact us with any queries.



21/10/2022, 17:09 Questionnaire for Study

Questionnaire for Study

Can sound produced from acoustic modelling be distinguished from sound produced from real world measurements in reverberant virtual spaces?

The goal of this study is to compare the perceived validity of sound fabricated using acoustic modelling techniques and sound produced using real world measurements, in varying virtual spaces.

Participation in this study is entirely voluntary, anonymous, and confidential (only seen by myself and my supervisor). You can decide whether or not to answer any of the questions and you can withdraw at anytime from the experiment without giving a reason.

The data collected will most likely be used in my dissertation and potentially other publications. The data will be stored for the duration of my project (another year), after which it will be destroyed.

When using virtual reality systems, some people may experience some degree of nausea. If at any time you wish to stop taking part in the study due to this or any other reason, please just stop.

If you have further queries after the study, feel free to drop me an email.

* Required

Instructions

Please refer back to the 'Instructions sheet' if you get stuck at any time. In short, when you enter the VR scenario, you will start in a hub world. Read the information there, and then you can proceed to scenario A. Make sure to select one of the two acoustic models before you play the sound sources. If this makes no sense, read the sheet!

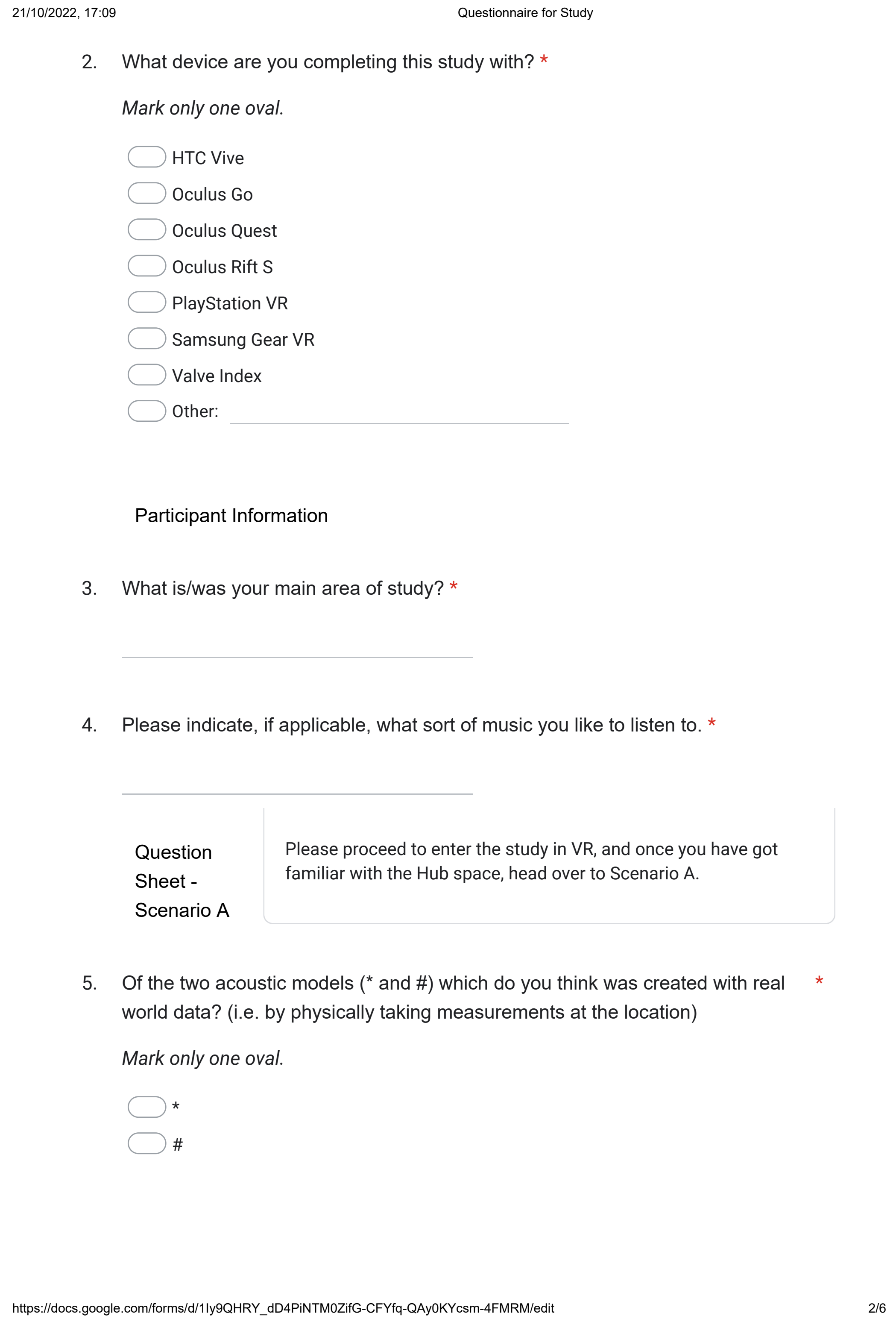
Study Code

1. STUDY CODE (optional). Please type a memorable code that you can quote if you would like to be removed from the study. We suggest using your initials followed by your birth date and month to make it easy to remember, but you are welcome to use whatever. Please write it down for future reference. If you do not leave a code, you will be unable to withdraw your data at a later date.

____

Headset

https://docs.google.com/forms/d/1Iy9QHRY_dD4PiNTM0ZifG-CFYfq-QAy0KYcsm-4FMRM/edit 1/6



2. What device are you completing this study with? *

Mark only one oval.

○ HTC Vive
○ Oculus Go
○ Oculus Quest
○ Oculus Rift S
○ PlayStation VR
○ Samsung Gear VR
○ Valve Index
○ Other: ______________________

Participant Information

3. What is/was your main area of study? *

____

4. Please indicate, if applicable, what sort of music you like to listen to. *

____

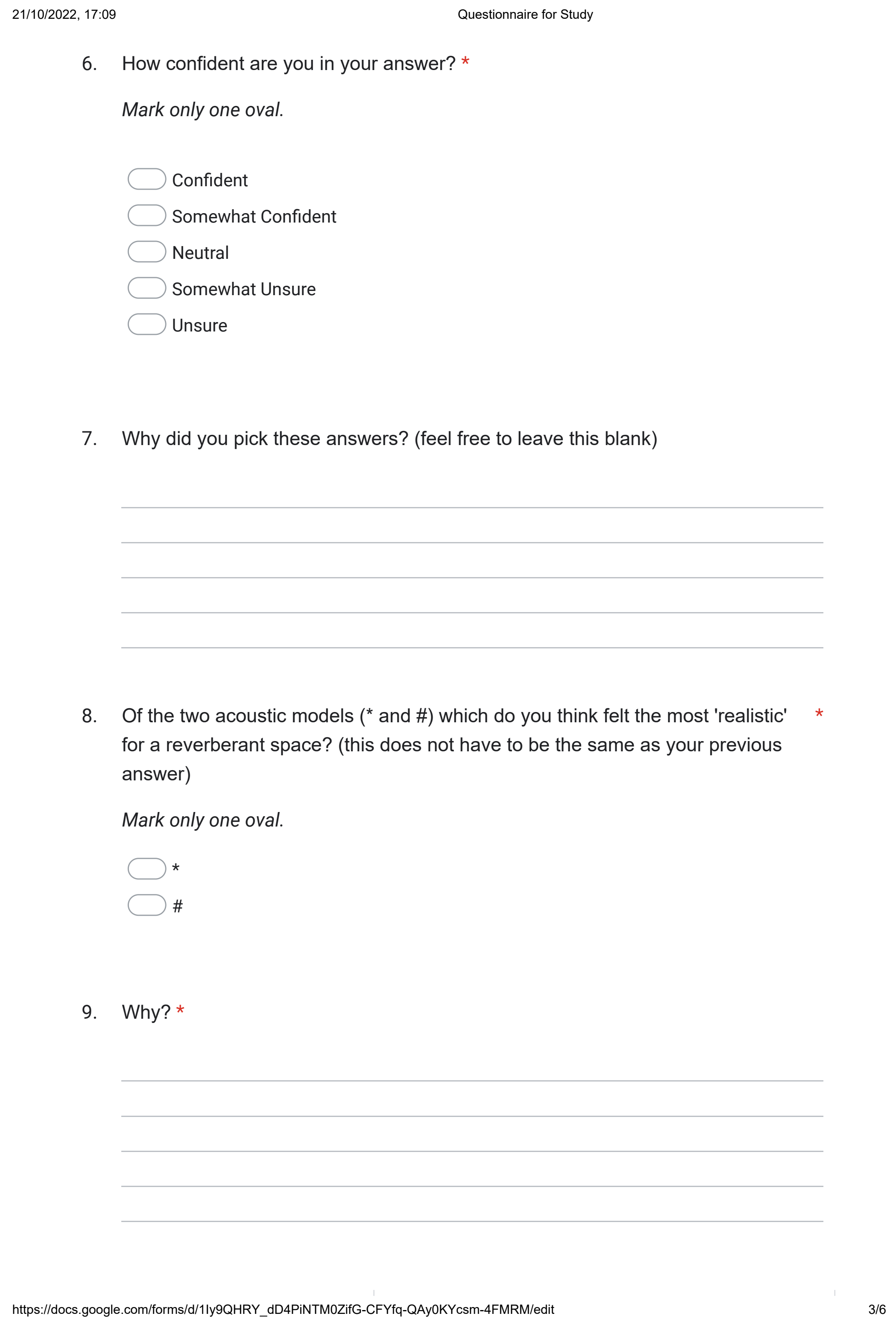
Question Sheet - Scenario A

Please proceed to enter the study in VR, and once you have got familiar with the Hub space, head over to Scenario A.

5. Of the two acoustic models (* and #) which do you think was created with real world data? (i.e. by physically taking measurements at the location) *

Mark only one oval.

○ *
○ #



6. How confident are you in your answer? *

Mark only one oval.

○ Confident
○ Somewhat Confident
○ Neutral
○ Somewhat Unsure
○ Unsure

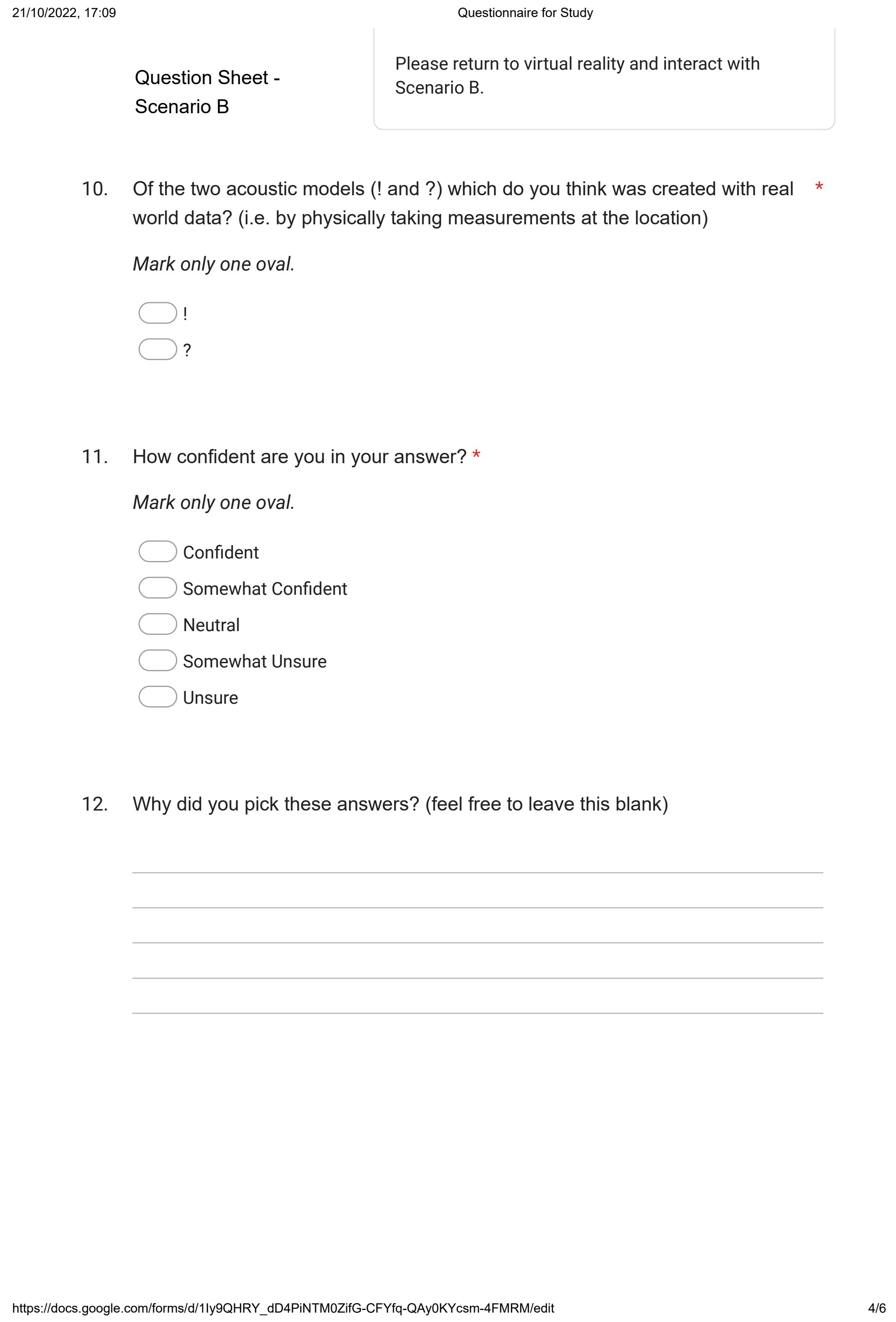
7. Why did you pick these answers? (feel free to leave this blank)

8. Of the two acoustic models (* and #) which do you think felt the most 'realistic' for a reverberant space? (this does not have to be the same as your previous answer) *

Mark only one oval.

○ *
○ #

9. Why? *



Question Sheet - Scenario B

Please return to virtual reality and interact with Scenario B.

10. Of the two acoustic models (! and ?) which do you think was created with real world data? (i.e. by physically taking measurements at the location) *

Mark only one oval.

○ !

○ ?

11. How confident are you in your answer? *

Mark only one oval.

○ Confident

○ Somewhat Confident

○ Neutral

○ Somewhat Unsure

○ Unsure

12. Why did you pick these answers? (feel free to leave this blank)

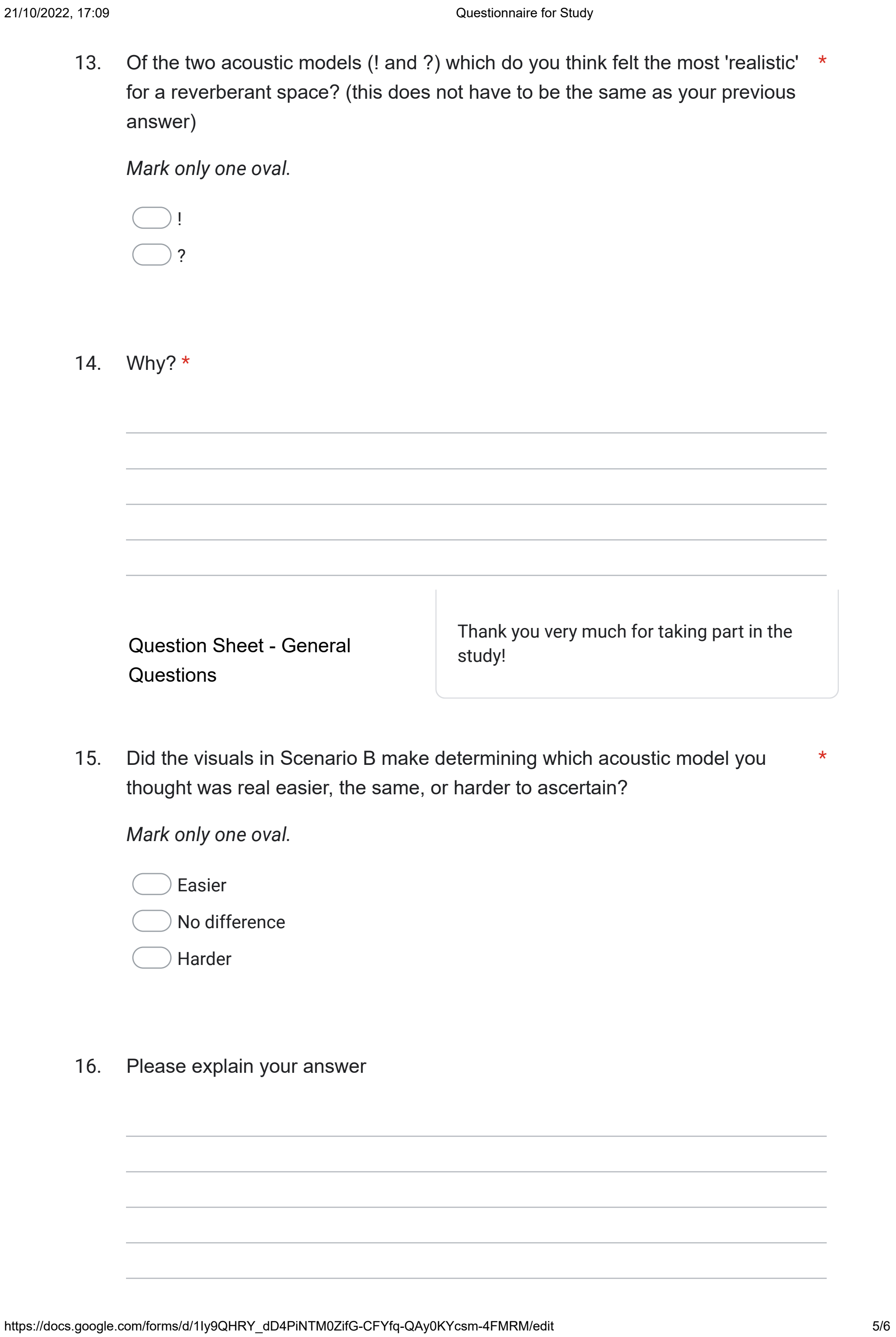
____

____

____

____

____



21/10/2022, 17:09 Questionnaire for Study

13. Of the two acoustic models (! and ?) which do you think felt the most 'realistic' for a reverberant space? (this does not have to be the same as your previous answer) *

Mark only one oval.

○ !

○ ?

14. Why? *

Question Sheet - General Questions

Thank you very much for taking part in the study!

15. Did the visuals in Scenario B make determining which acoustic model you thought was real easier, the same, or harder to ascertain? *

Mark only one oval.

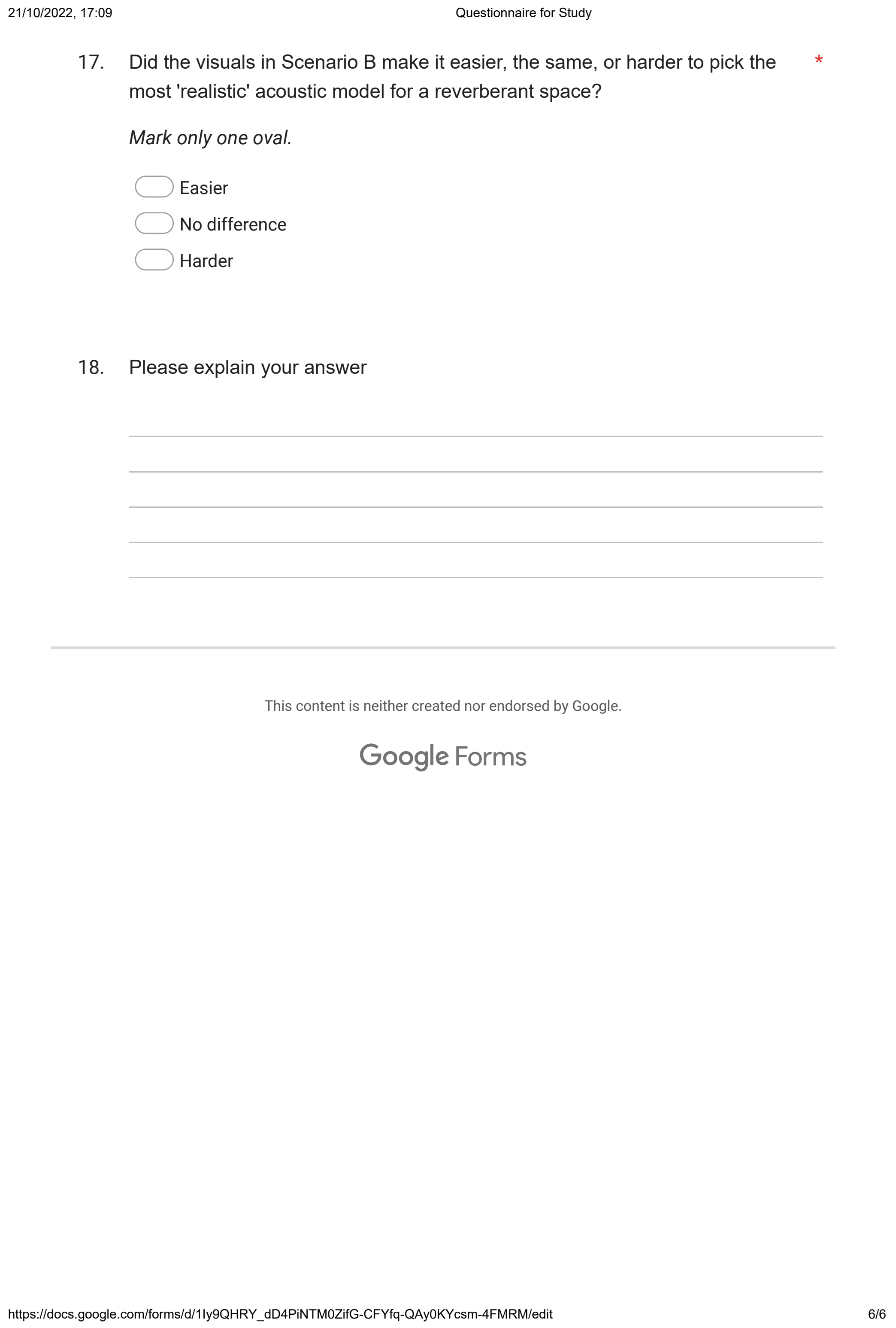
○ Easier

○ No difference

○ Harder

16. Please explain your answer

https://docs.google.com/forms/d/1Iy9QHRY_dD4PiNTM0ZifG-CFYfq-QAy0KYcsm-4FMRM/edit 5/6



17. Did the visuals in Scenario B make it easier, the same, or harder to pick the most 'realistic' acoustic model for a reverberant space? *

Mark only one oval.

○ Easier

○ No difference

○ Harder

18. Please explain your answer

This content is neither created nor endorsed by Google.

Google Forms

Appendix C

Instructions and Questionnaire for Study 3 (Chapter 5)

Information Sheet

Electronic Engineering,

Royal Holloway, University of London

Name of study: Some sound cues are more important for immersivity than others for reverberant spaces.

Name of Researcher: Florence Roberts (PhD Student) Email: zavc374@live.rhul.ac.uk

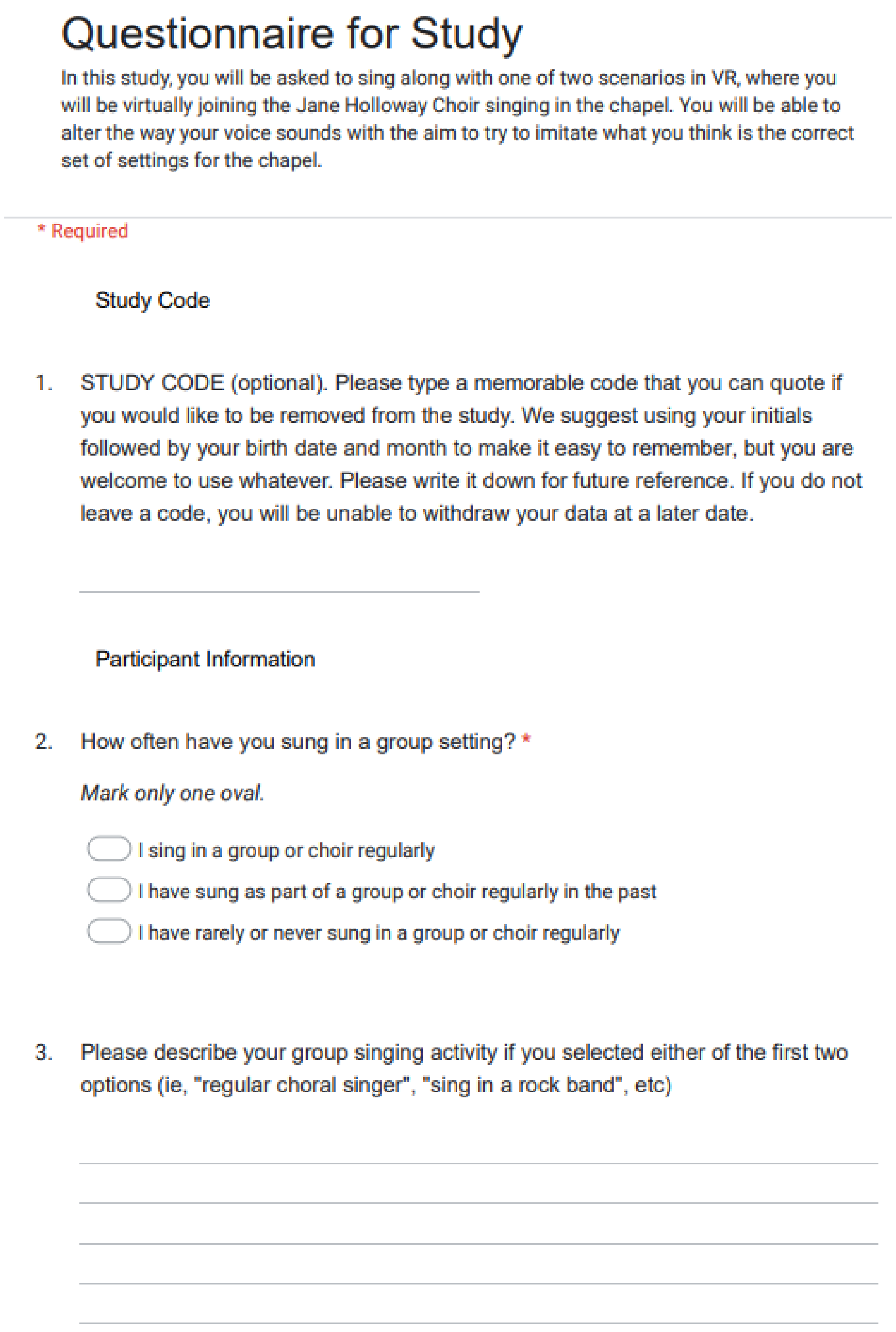
In this study, you will be asked to sing along with one of two scenarios in VR, where you will be virtually joining the Jane Holloway Choir singing in the chapel. You will be able to alter the way you hear your voice with the aim to try to imitate what you think is the correct set of settings for the chapel.

Participation in this study is entirely voluntary, anonymous, and confidential (only seen by myself and my supervisor). You can decide whether or not to answer any of the questions and you can withdraw at any time from the experiment without giving a reason.

The data collected will most likely be used in my dissertation and potentially other publications. The data will be stored for the duration of my project, after which it will be destroyed. For this duration, the consent form and your answers will be stored on my computer.

If you have any questions or complaints during the study, just let me know. If you have further queries after the study, feel free to drop me an email.

If you are happy to participate in this study, please tell the researcher and they will provide you with the consent form.



Questionnaire for Study

In this study, you will be asked to sing along with one of two scenarios in VR, where you will be virtually joining the Jane Holloway Choir singing in the chapel. You will be able to alter the way your voice sounds with the aim to try to imitate what you think is the correct set of settings for the chapel.

* Required

Study Code

1. STUDY CODE (optional). Please type a memorable code that you can quote if you would like to be removed from the study. We suggest using your initials followed by your birth date and month to make it easy to remember, but you are welcome to use whatever. Please write it down for future reference. If you do not leave a code, you will be unable to withdraw your data at a later date.

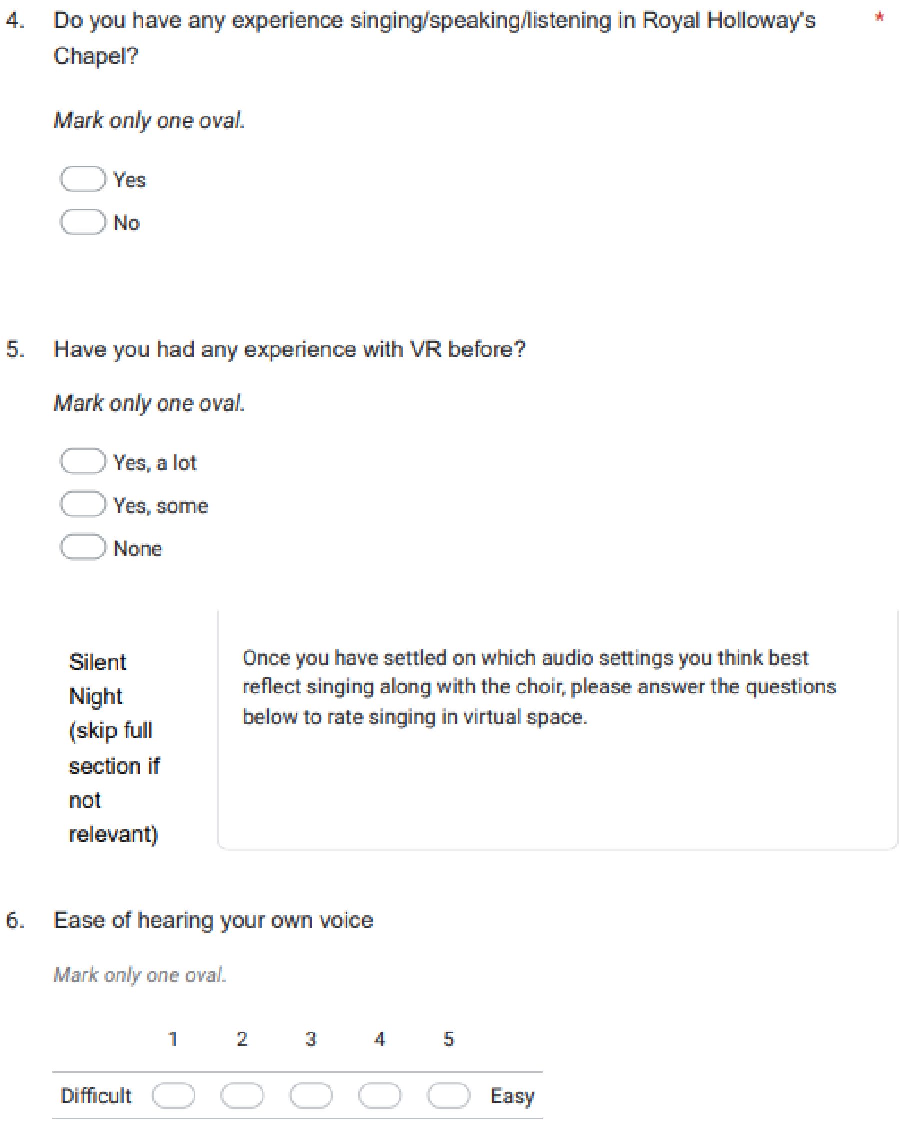
Participant Information

2. How often have you sung in a group setting? *

Mark only one oval.

○ I sing in a group or choir regularly
○ I have sung as part of a group or choir regularly in the past
○ I have rarely or never sung in a group or choir regularly

3. Please describe your group singing activity if you selected either of the first two options (ie, "regular choral singer", "sing in a rock band", etc)



4. Do you have any experience singing/speaking/listening in Royal Holloway's Chapel? *

Mark only one oval.

○ Yes
○ No

5. Have you had any experience with VR before?

Mark only one oval.

○ Yes, a lot
○ Yes, some
○ None

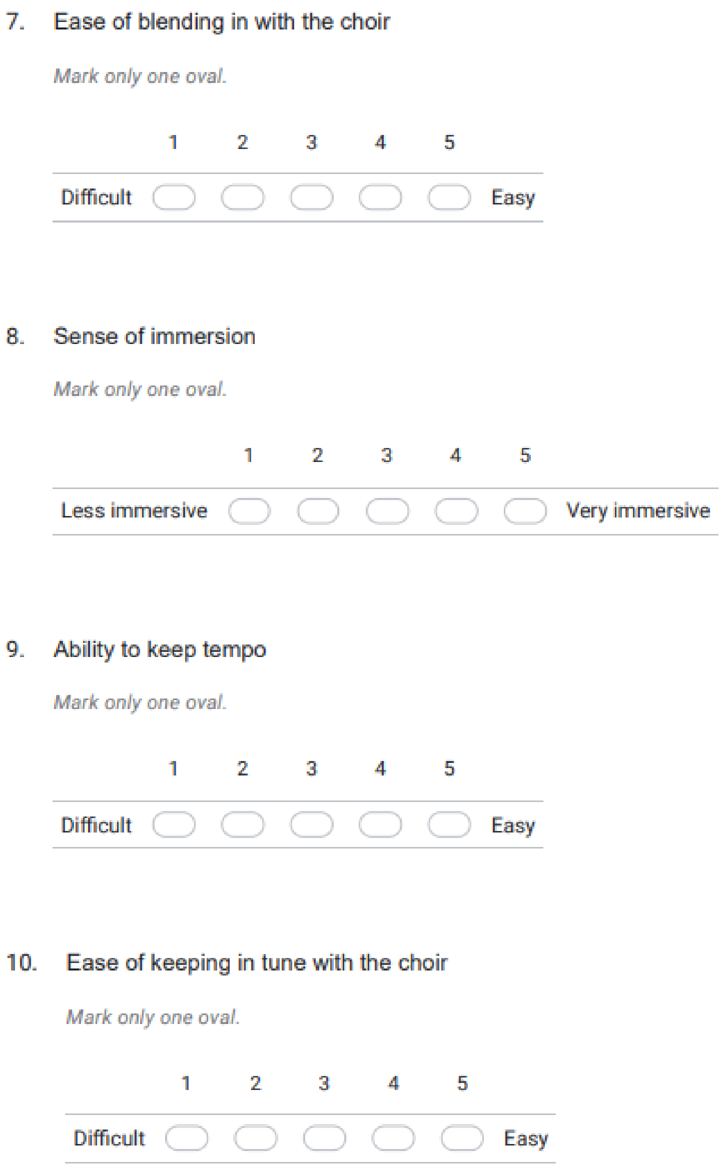
Silent Night (skip full section if not relevant)

Once you have settled on which audio settings you think best reflect singing along with the choir, please answer the questions below to rate singing in virtual space.

6. Ease of hearing your own voice

Mark only one oval.

	1	2	3	4	5	
Difficult	○	○	○	○	○	Easy



7. Ease of blending in with the choir

Mark only one oval.

	1	2	3	4	5	
Difficult	○	○	○	○	○	Easy

8. Sense of immersion

Mark only one oval.

	1	2	3	4	5	
Less immersive	○	○	○	○	○	Very immersive

9. Ability to keep tempo

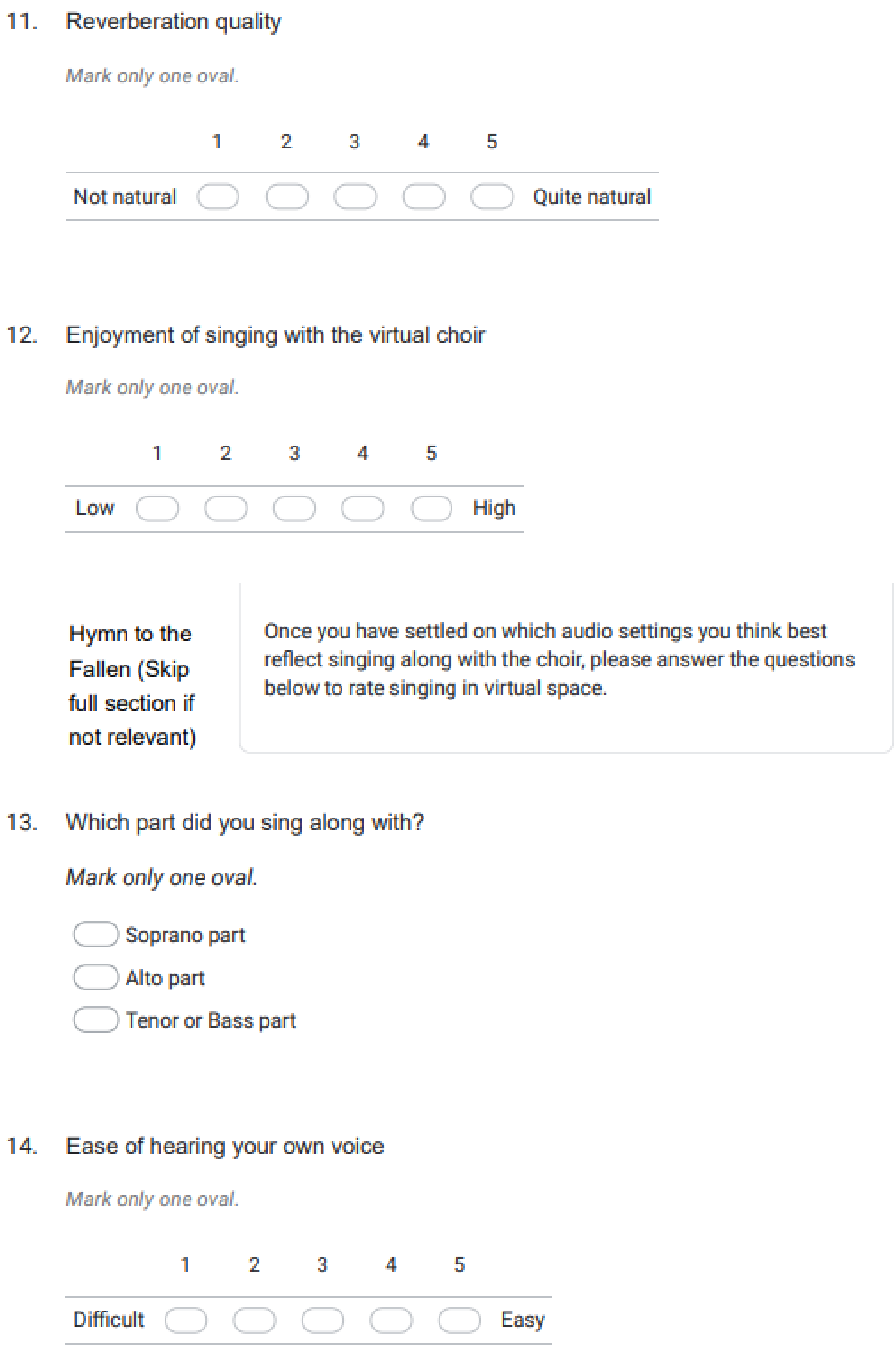
Mark only one oval.

	1	2	3	4	5	
Difficult	○	○	○	○	○	Easy

10. Ease of keeping in tune with the choir

Mark only one oval.

	1	2	3	4	5	
Difficult	○	○	○	○	○	Easy



11. Reverberation quality

Mark only one oval.

	1	2	3	4	5	
Not natural	○	○	○	○	○	Quite natural

12. Enjoyment of singing with the virtual choir

Mark only one oval.

	1	2	3	4	5	
Low	○	○	○	○	○	High

Hymn to the Fallen (Skip full section if not relevant)

Once you have settled on which audio settings you think best reflect singing along with the choir, please answer the questions below to rate singing in virtual space.

13. Which part did you sing along with?

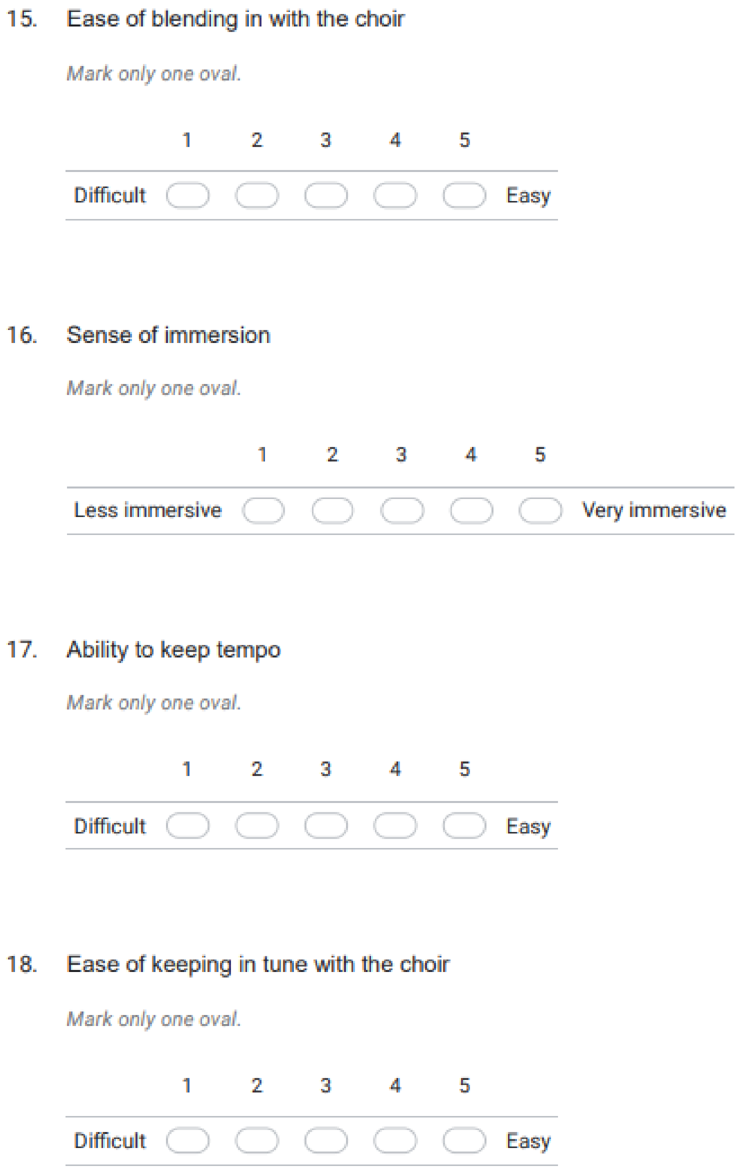
Mark only one oval.

○ Soprano part
○ Alto part
○ Tenor or Bass part

14. Ease of hearing your own voice

Mark only one oval.

	1	2	3	4	5	
Difficult	○	○	○	○	○	Easy



15. Ease of blending in with the choir

Mark only one oval.

	1	2	3	4	5	
Difficult	○	○	○	○	○	Easy

16. Sense of immersion

Mark only one oval.

	1	2	3	4	5	
Less immersive	○	○	○	○	○	Very immersive

17. Ability to keep tempo

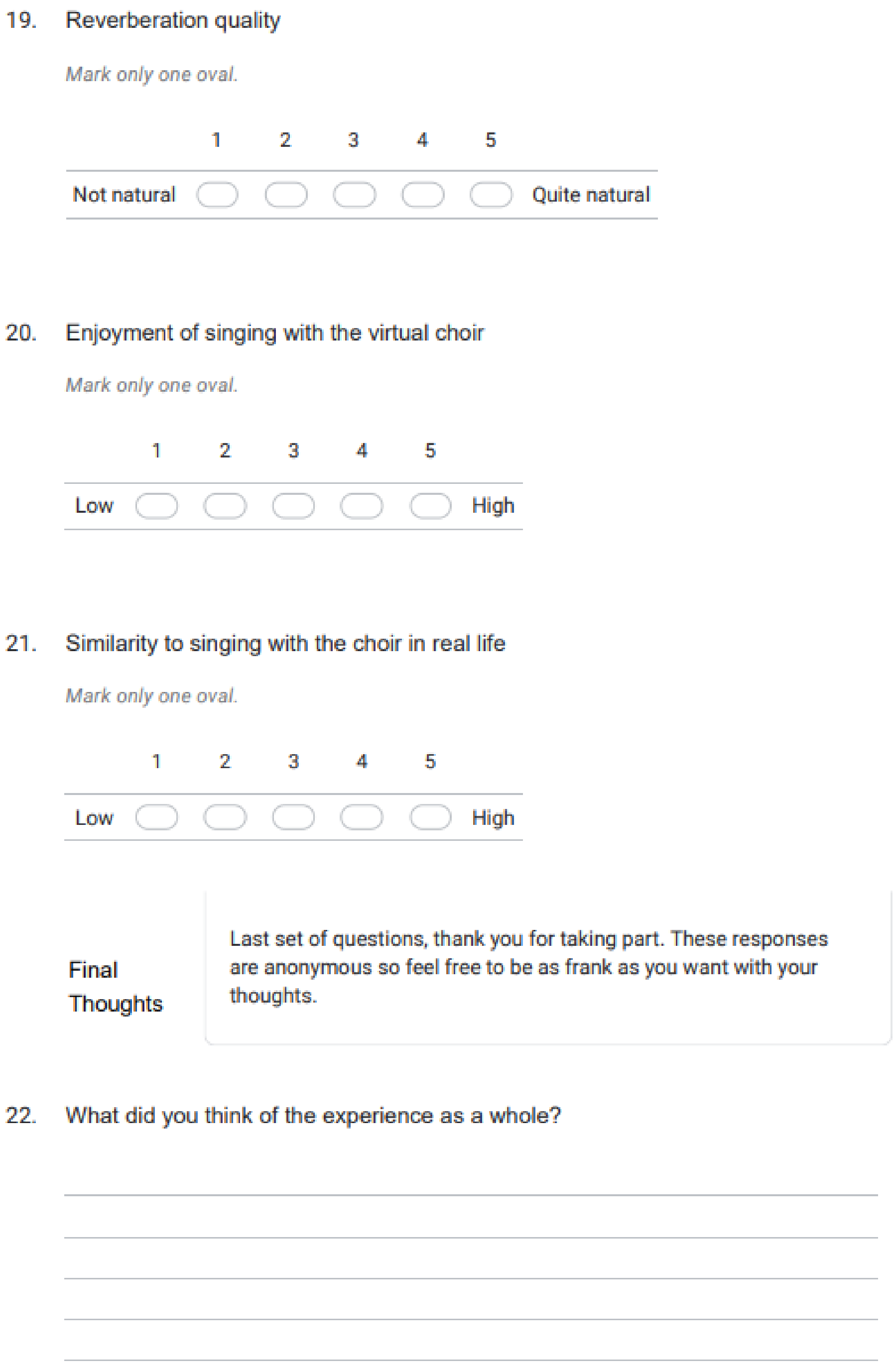
Mark only one oval.

	1	2	3	4	5	
Difficult	○	○	○	○	○	Easy

18. Ease of keeping in tune with the choir

Mark only one oval.

	1	2	3	4	5	
Difficult	○	○	○	○	○	Easy



19. Reverberation quality

Mark only one oval.

	1	2	3	4	5	
Not natural	○	○	○	○	○	Quite natural

20. Enjoyment of singing with the virtual choir

Mark only one oval.

	1	2	3	4	5	
Low	○	○	○	○	○	High

21. Similarity to singing with the choir in real life

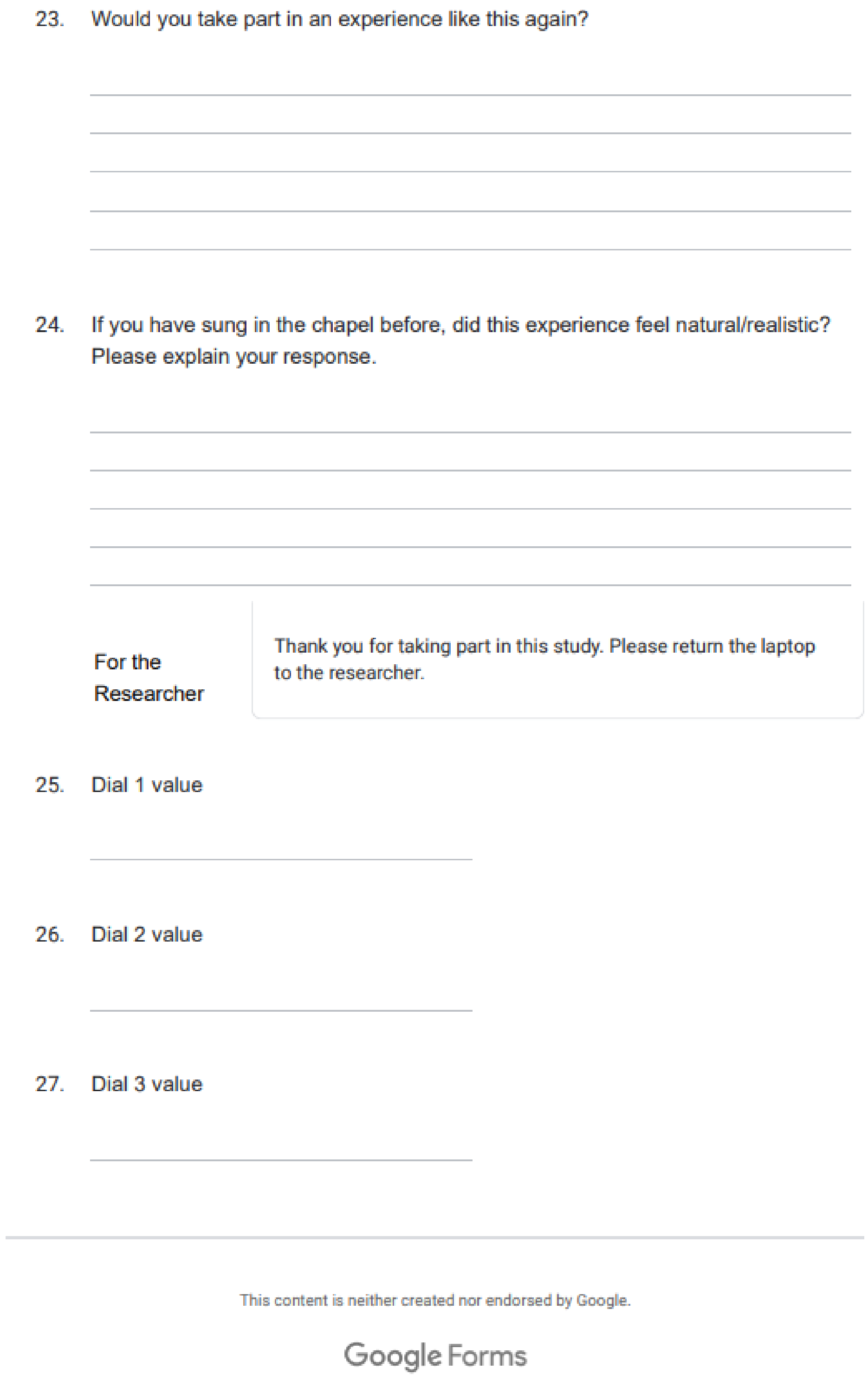
Mark only one oval.

	1	2	3	4	5	
Low	○	○	○	○	○	High

Final Thoughts

Last set of questions, thank you for taking part. These responses are anonymous so feel free to be as frank as you want with your thoughts.

22. What did you think of the experience as a whole?



23. Would you take part in an experience like this again?

24. If you have sung in the chapel before, did this experience feel natural/realistic? Please explain your response.

For the Researcher

Thank you for taking part in this study. Please return the laptop to the researcher.

25. Dial 1 value

26. Dial 2 value

27. Dial 3 value

This content is neither created nor endorsed by Google.

Bibliography

[1] N. O-larnnithipong, A. B. Barreto, and F. Abyarjoo, “Impact of binaural 3D sound on navigation within a virtual environment,” 2018.

[2] K. Farrar, “Soundfield microphone,” *Wireless World*, vol. 85, no. 1526, pp. 48–50, 1979.

[3] Google, “Fundamental concepts.” https://resonance-audio.github.io/resonance-audio/discover/concepts.html, 2018. Accessed: 21/03/2021.

[4] Resonance Audio, “Fundamental concepts.” https://developers.google.com/resonance-audio/discover/concepts, 2017. Accessed: 21/08/2019.

[5] L. Álvarez-Morales, J. Molina, S. Girón, A. Alonso, P. Bustamante, and Á. Álvarez-Corbacho, “Virtual reality in church acoustics: Visual and acoustic experience in the cathedral of seville, spain,” 07 2017.

[6] G. Kearney, H. Daffern, L. Thresh, H. Omodudu, C. Armstrong, and J. Brereton, “Design of an interactive virtual reality system for ensemble singing,” in *Proceedings of the Interactive Audio Systems Symposium*, 2016.

[7] MH Acoustics, “Anker Soundcore 2 Speaker Review.” https://www.rtings.com/speaker/reviews/anker/soundcore-2, 2021.

[8] G. Regal, R. Schatz, J. Schrammel, and S. Suette, “Vrate: A Unity3D asset for integrating subjective assessment questionnaires in virtual environments,” in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3, 2018.

[9] M. Feick, N. Kleer, A. Tang, and A. Krüger, “The virtual reality questionnaire toolkit,” in *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST ’20 Adjunct, (New York, NY, USA), p. 68–69, Association for Computing Machinery, 2020.

[10] Bela, “Bela IDE.” http://bela.local/. Accessed: 21/03/2021.

[11] R. M. Schafer, *The Soundscape: Our Sonic Environment and the Tuning of the World*. Inner Traditions Bear and Company, 1993.

[12] C. Ball, K.-T. Huang, and J. Francis, “Virtual reality adoption during the covid-19 pandemic: A uses and gratifications perspective,” *Telematics and Informatics*, vol. 65, p. 101728, 2021.

[13] S. Agrawal, A. Simon, S. Bech, K. Bærentsen, and S. Forchhammer, “Defining immersion: Literature review and implications for research on immersive audiovisual experiences,” in *Audio Engineering Society Convention 147*, Oct 2019.

[14] H. Lee, “A conceptual model of immersive experience in extended reality,” *PsyArXiv*, 09 2020.

[15] C. Eaton, *Quantifying Factors of Auditory Immersion for Virtual Reality*. PhD thesis, University of Huddersfield, 2020.

[16] Cambridge Dictionary, “Immersion.” https://dictionary.cambridge.org/dictionary/english/immersion, year=2023.

[17] Oxford Dictionary, “Immersive.” https://en.oxforddictionaries.com/definition/immersive, year=2019.

[18] M.-L. Ryan, “Narrative as virtual reality: Immersion and interactivity in literature an electronic media,” *South Atlantic Review*, vol. 67, 01 2004.

[19] M. Lombard and T. Ditton, “At the Heart of It All: The Concept of Presence,” *Journal of Computer-Mediated Communication*, vol. 3, 09 1997.

[20] D. Arsenault, “Dark waters: Spotlight on immersion,” 01 2005.

[21] F. Biocca and B. Delaney, “Immersive virtual reality technology,” 1995.

[22] A. McMahan, “Immersion, engagement, and presence: A method for analyzing 3-d video games,” 2013.

[23] L. Ermi and F. Mäyrä, “Fundamental components of the gameplay experience: Analysing immersion.,” 01 2005.

[24] E. Adams and A. Rollings, *Game design and development: Fundamentals of game design*. New Jersey, Pearson Prentice Hall, 2007.

[25] J. Murray, *Hamlet on the Holodeck: The Future of Narrative in Cyberspace* . Cambridge, MA: The MIT Press, 1997.

[26] A. Witmer and M. Slater, “Measuring presence: A response to the witmer and singer presence questionnaire,” *Presence (Camb.)*, vol. 8, 12 1999.

[27] M. Slater, “A note on presence terminology,” *Presence Connect*, vol. 3, 01 2003.

[28] E. Ridgway, "Ear speakers - research, design, and evolution," *Steam*, 2019.

[29] Y. Huang, J. Chen, and J. Benesty, "Immersive audio schemes," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 20–32, 2011.

[30] M. Geronazzo, *Immersive Auralization Using Headphones*, pp. 1–5. Cham: Springer International Publishing, 2018.

[31] M. Gorzel, A. Allen, I. Kelly, A. Gungormusler, J. Kammerl, H. Yeh, and F. Boland, "Efficient encoding and decoding of binaural sound with resonance audio," 03 2019.

[32] S. Kim and W. Howie, "Influence of the listening environment on recognition of immersive reproduction of orchestral music sound scenes," *Journal of the Audio Engineering Society*, vol. 69, pp. 834–848, 11 2021.

[33] D. Satongar, *Simulation And Analysis Of Spatial Audio Reproduction And Listening Area Effects*. PhD thesis, University of Salford, 2016.

[34] R. Stevens and D. Raybould, "The reality paradox: Authenticity, fidelity and the real in battlefield 4," *The Soundtrack*, vol. 8, pp. 57–75, 10 2015.

[35] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *Journal of the Audio Engineering Society*, vol. 44, pp. 451–469, june 1996.

[36] F. Rumsey, "Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm," *Journal of the Audio Engineering Society*, vol. 50, no. 9, p. 651–666, 2002.

[37] A. Wolfe, "Palmer Luckey: Making virtual reality a reality," *The Wall Street Journal.*

[38] W. E. Carlson, *Computer Graphics and Computer Animation: A Retrospective Overview*. The Ohio State University, 2005.

[39] K. Williams, "The virtual arena – blast from the past: The VR-1," *Good Morning Web 3*, 2020.

[40] S. L. Kent, *The Ultimate History of Video Games: The Story Behind the Craze that Touched Our Lives and Changed the World*. Random House International, 2002.

[41] Gamereactor, "E3 12: John Carmack's VR Presentation." https://www.youtube.com/watch?v=kw-DlWwlXHo&ab_channel=Gamereactor, 2012.

[42] D. Heaney and UploadVR, "How virtual reality positional tracking works." https://venturebeat.com/2019/05/05/how-virtual-reality-positional-tracking-works/3/, 2019.

[43] L. Gu, D. Cheng, and Y. Wang, "Design of an immersive head mounted display with coaxial catadioptric optics," in *Digital Optics for Immersive Displays* (B. C. Kress, W. Osten, and H. Stolle, eds.), vol. 10676, pp. 353–358, International Society for Optics and Photonics, SPIE, 2018.

[44] Steam, "Valve Index: Controllers." https://www.valvesoftware.com/en/index/controllers.

[45] J. Linneman and A. Battaglia, "Half-Life: Alyx tech analysis - a VR masterpiece that must be experienced," *Eurogamer*.

[46] J. Feltham, "Breath Tech Is A VR Puzzle Game That Uses Your Breath," *UploadVR*.

[47] SensoryCo, "4D Theatre Effects and Scenting." https://sensoryco4d.com/4d-theatre-effects-scenting/, 2020. Accessed: 08/12/2020.

[48] H. Kim, L. Hernaggi, P. J. Jackson, and A. Hilton, "Immersive spatial audio reproduction for VR/AR using room acoustic modelling from 360° images," *26th IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2019 - Proceedings*, pp. 120–126, 2019.

[49] W. Zhao, D. Nister, and S. Hsu, "Alignment of continuous video onto 3d point clouds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1305–1318, 2005.

[50] U. Neumann and S. You, "Modeling and video projection for augmented virtual environments," Sept. 1 2009. US Patent 7,583,275.

[51] S. Nebiker, S. Bleisch, and M. Christen, "Rich point clouds in virtual globes–a new paradigm in city modeling?," *Computers, Environment and Urban Systems*, vol. 34, no. 6, pp. 508–517, 2010.

[52] G. Bruder, F. Steinicke, and A. Nüchter, "Poster: Immersive point cloud virtual environments," in *2014 IEEE symposium on 3D user interfaces (3DUI)*, pp. 161–162, IEEE, 2014.

[53] G. Bui, B. Morago, T. Le, K. Karsch, Z. Lu, and Y. Duan, "Integrating videos with lidar scans for virtual reality," in *2016 IEEE Virtual Reality (VR)*, pp. 161–162, 2016.

[54] GoPro, "GoPro Fusion 360-degree Camera." https://gopro.com/en/gb/shop/cameras/fusion/CHDHZ-101-master.html, 2017.

[55] M. Binelli, D. Pinardi, A. Farina, and T. Nili, "Individualized HRTF for playing VR videos with Ambisonics spatial audio on HMDs," *Journal of the Audio Engineering Society*, august 2018.

[56] C. Eaton and H. Lee, “Quantifying factors of auditory immersion in virtual reality,” in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Mar 2019.

[57] A. McArthur, M. Sandler, and R. Stewart, “Accuracy of perceived distance in VR using verbal descriptors,” in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Mar 2019.

[58] O. Rummukainen, T. Robotham, S. Schlecht, A. Plinge, J. Herre, and E. Habets, “Audio quality evaluation in virtual reality: multiple stimulus ranking with behavior tracking,” in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, 08 2018.

[59] G. C. Stecker, T. M. Moore, M. Folkerts, D. Zotkin, and R. Duraiswami, “Toward objective measures of auditory co-immersion in virtual and augmented reality,” in *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*, Aug 2018.

[60] A. Mcarthur, M. Sandler, and R. Stewart, “Perception of mismatched auditory distance - cinematic VR,” *Journal of the Audio Engineering Society*, august 2018.

[61] R. Selfridge, J. Cook, K. McAlpine, and M. Newton, “Creating historic spaces in virtual reality using off-the-shelf audio plugins,” in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Mar 2019.

[62] M. Slater, “Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments,” *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 364, pp. 3549–57, 12 2009.

[63] J. Blauert, *Spatial hearing: The psychophysics of human sound localization.* MIT Press, 1997.

[64] B. Grothe, M. Pecka, and D. McAlpine, "Mechanisms of sound localization in mammals," *Physiological Reviews*, vol. 90, no. 3, pp. 983–1012, 2010. PMID: 20664077.

[65] E. B. Goldstein, *Sensation and Perception (8th Edition).* Wadsworth Publishing, 2010.

[66] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-Related Transfer Functions of human subjects," *journal of the audio engineering society*, vol. 43, pp. 300–321, may 1995.

[67] S. Li and J. Peissig, "Measurement of Head-Related Transfer Functions: A review," *Applied Sciences*, vol. 10, p. 5014, 07 2020.

[68] J. Watkinson, *The Art of Sound Reproduction.* Oxford: Taylor & Francis, 1998.

[69] J. Steinberg and W. Swow, "Auditory perspective-physical factors," *Transactions of the American Institute of Electrical Engineers*, vol. 53, pp. 12–17, 1934.

[70] Audio Engineering Society, *An anthology of reprinted articles on stereophonic techniques.* 1986.

[71] W. Gardner, *3-D Audio Using Loudspeakers.* PhD thesis, Massachusetts Institute of Technology, 1998.

[72] F. Rumsey, *Spatial Audio.* Oxford: Focal Press, 2001.

[73] H. F. Olson, *Modern Sound Reproduction.* New York: Van Nostrand Reinhold Inc., 1972.

[74] E. Torick, “Highlights in the History of Multichannel Sound,” *AES: Journal of the Audio Engineering Society*, vol. 46, no. 1-2, pp. 27–31, 1998.

[75] P. J. Philipson, J. Hirst, and S. Woollard, “Investigation into a method for predicting the perceived azimuth position of a virtual image,” *Journal of the Audio Engineering Society*, april 2002.

[76] J. D. Rees-Jones, J. S. Brereton, and D. T. Murphy, “Spatial audio quality and user preference of listening systems in video games,” pp. 223–230, 2015.

[77] 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, 2022.

[78] M. A. Gerzon, “Periphony: With-height sound reproduction,” *Journal of the Audio Engineering Society*, vol. 21, pp. 2–10, Feb 1973.

[79] P. Lecomte, P. A. Gauthier, C. Langrenne, A. Garcia, and A. Berry, “On the use of a lebedev grid for ambisonics,” *139th Audio Engineering Society International Convention, AES 2015*, pp. 1–12, 2015.

[80] Zoom, “Zoom H3-VR: Monitoring and Playback.” https://ambisonics.de/vr-audio-encoding-decoding/zoom-h3-vr-monitoring-and-playback, 2019.

[81] M. A. Gerzon and G. J. Barton, “Ambisonic decoders for HDTV,” *92nd Convention of the Audio Engineering Society*, p. Preprint 3345, 1992.

[82] M. Kronlachner and F. Zotter, “Spatial transformations for the enhancement of Ambisonic recordings,” *2nd International Conference on Spatial Audio*, no. 2, pp. 1–5, 2014.

[83] J.-M. Jot, V. Larcher, and J.-M. Pernaux, "A comparative study of 3-D audio encoding and rendering techniques," in *AES 16th International Conference*, Mar 1999.

[84] B. Wiggins, *An investigation into the real-time manipulation and control of three-dimensional sound fields.* PhD thesis, University of Derby, 2004.

[85] A. Farina, R. Glasgal, E. Armelloni, and A. Torger, "Ambiophonic principles for the recording and reproduction of surround sound for music," *Journal of the Audio Engineering Society*, june 2001.

[86] M. Grimshaw and T. Garner, "Defining Sound," in *Sonic Virtuality: Sound as Emergent Perception*, Oxford University Press, 08 2015.

[87] V. Michael, *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality.* Springer, 2011.

[88] M. Barron and T. J. Foulkes, *Auditorium Acoustics and Architectural Design* , vol. 96. 1994.

[89] M. Kleiner, B.-I. Dalënback, and P. Svensson, "Auralization-an overview," *Journal of Endocrinological Investigation*, vol. 41, no. 11, pp. 861–875, 1993.

[90] H. Lehnert and J. Blauert, "Principles of binaural room simulation," *Applied Acoustics*, vol. 36, no. 3, pp. 259–291, 1992.

[91] Acoustics Engineering, "Measuring Impulse Responses Using Dirac," *Acoustics Engineering*, no. August, pp. 1–30, 2007.

[92] J. S. Brereton, *Singing in Space(s): Singing performance in real and virtual acoustic environments—Singers' evaluation, performance analysis and listeners' perception.* PhD thesis, University of York, Electronics, 2014.

[93] H. Mi, G. Kearney, and H. Daffern, “Impact thresholds of parameters of binaural room impulse responses (brirs) on perceptual reverberation,” *Applied Sciences*, vol. 12, no. 6, 2022.

[94] D. Howard and J. Angus, *Acoustics and Psychoacoustics.* Routledge: Abingdon-on-Thames, UK, 2013.

[95] S. Müller and P. Massarani, “Transfer-Function Measurement with Sweeps,” *AES: Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 443–471, 2001.

[96] G. B. Stan, J. J. Embrechts, and D. Archambeau, “Comparison of different impulse response measurement techniques,” *AES: Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249–262, 2002.

[97] A. Farina, “Simultaneous measurement of impulse response and distortion with a swept-sine technique,” in *Audio Engineering Society Convention 108*, Feb 2000.

[98] L. Remaggi, H. Kim, P. J. B. Jackson, and A. Hilton, “Reproducing real world acoustics in virtual reality using spherical cameras,” in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Mar 2019.

[99] V. Välimäki, J. Parker, L. Savioja, J. O. Smith, and J. Abel, “More than 50 years of artificial reverberation,” in *AES 60th International Conference: DREAMS*, Audio Engineering Society, 2016.

[100] V. Välimäki and J. D. Reiss, “All about audio equalization: Solutions and frontiers,” *Applied Sciences*, vol. 6, no. 5, 2016.

[101] C. Steinmetz, “Final project: Zero-latency convolution on bela.” https://github.com/csteinmetz1/bela-zlc, 2021. Accessed: 26/04/2022.

[102] C. D. McGillem and G. R. Cooper, *Continuous and discrete signal and system analysis.* Holt Rinehart and Winston, 1984.

[103] F. Wefers, *Partitioned convolution algorithms for real-time auralization*, vol. 20. Logos Verlag Berlin GmbH, 2015.

[104] J. Hurchalla, "A time distributed fft for efficient low latency convolution," in *129th Audio Engineering Society Convention*, Audio Engineering Society, 2010.

[105] U. Svensson, "Modelling acoustic spaces for audio virtual reality," in *Proceedings of the 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, 01 2002.

[106] N. Raghuvanshi, N. Galoppo, and M. C. Lin, "Accelerated wave-based acoustics simulation," in *Proceedings of the 2008 ACM Symposium on Solid and Physical Modeling*, SPM '08, (New York, NY, USA), p. 91–102, Association for Computing Machinery, 2008.

[107] B. Kapralos, M. Jenkin, and E. Milios, "Sonel mapping: A probabilistic acoustical modeling method," *Building Acoustics*, vol. 15, 12 2008.

[108] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015.

[109] M. Vorländer, D. Schröder, S. Pelzer, and F. Wefers, "Virtual reality for architectural acoustics," *Journal of Building Performance Simulation*, vol. 8, no. 1, pp. 15–25, 2015.

[110] F. Pind, C.-H. Jeong, H. S. Llopis, K. Kosikowski, and J. Strømann-Andersen, "Acoustic Virtual Reality – Methods and challenges," *Baltic-Nordic Acoustics Meeting*, no. April 2018, pp. 1–11, 2018.

[111] T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West, "A beam tracing approach to acoustic modeling for interactive virtual environments," in *25th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 21–32, 1998.

[112] H. Kuttruff, *Room Acoustics*. CRC Press, 5th ed., 2009.

[113] H. Lee and B.-H. Lee, "An efficient algorithm for the image model technique," *Applied Acoustics*, vol. 24, no. 2, pp. 87–115, 1988.

[114] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 04 1979.

[115] A. Kulowski, "Algorithmic representation of the ray tracing technique," *Applied Acoustics*, vol. 18, no. 6, pp. 449–469, 1985.

[116] B. Miga and B. Ziołko, "Real-Time Acoustic Phenomena Modelling for Computer Games Audio Engine," *Archives of Acoustics*, vol. 40, no. 2, pp. 205–211, 2015.

[117] H. Lehnert, "Systematic errors of the ray-tracing algorithm," *Applied Acoustics*, vol. 38, no. 2, pp. 207–221, 1993.

[118] Google, "Developing with Resonance Audio." https://resonance-audio.github.io/resonance-audio/develop/overview.html, 2018. Accessed: 20/03/2021.

[119] A. Algargoosh, "Review of aspects that shape the aural experience in worship spaces," in *Proceedings of Meetings on Acoustics*, vol. 28, 09 2016.

[120] J. Rees-Jones and H. Daffern, "The hills are alive: Capturing and presenting an outdoor choral performance for virtual reality," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Mar 2019.

[121] MH Acoustics, "EM32 Eigenmike microphone array release notes (v17.0)." https://www.mhacoustics.com/sites/default/files/ReleaseNotes.pdf, 2003.

[122] S. Bertet, J. Daniel, L. Gros, E. Parizet, and O. Warusfel, "Investigation of the perceived spatial resolution of higher order ambisonics sound fields: A subjective evaluation involving virtual and real 3D microphones," in *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*, Mar 2007.

[123] M. Frank, F. Zotter, H. Wierstorf, and S. Spors, *Spatial Audio Rendering*, pp. 247–260. SPRINGER, 2014.

[124] E. Bates, G. Kearney, D. Furlong, and F. Boland, "Localization accuracy of advanced spatialisation techniques in small concert halls," *The Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 3069–3070, 2007.

[125] L. Reed, M. Harries, A. Hurr, and M. Knight, "Applied multichannel recording of a contemporary symphony orchestra for virtual reality," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Mar 2019.

[126] KanDaoVR, "KanDao Obsidian S/R." https://www.kandaovr.com/obsidian-s-r/, 2019.

[127] Facebook Audio 360, "Facebook 360 spatial workstation version 2.2.1." https://facebook360.fb.com/spatial-workstation/, 2019.

[128] S. Girón, M. Galindo, and T. Gómez-Gómez, “Assessment of the subjective perception of reverberation in spanish cathedrals,” *Building and Environment*, vol. 171, p. 106656, 2020.

[129] N. Kaplanis, S. Bech, T. Lokki, T. van Waterschoot, and S. Holdt Jensen, “Perception and preference of reverberation in small listening rooms for multi-loudspeaker reproduction,” *The Journal of the Acoustical Society of America*, vol. 146, pp. 3562–3576, 11 2019.

[130] M. Gölzer, H.; Kleinschmidt, “M. importance of early and late reflections for automatic speech recognition in reverberant environments,” *In Proceedings of the Elektronische Sprachsignalverarbeitung (ESSV)*, 2003.

[131] W. G. Gardner, *Reverberation Algorithms*, pp. 85–131. Boston, MA: Springer US, 2002.

[132] k. h. kuttruff, “Auralization of impulse responses modeled on the basis of ray-tracing results,” *journal of the audio engineering society*, vol. 41, pp. 876–880, november 1993.

[133] C. Shoard, “Walter Murch: Searching for the sound of the God particle,” *The Guardian*.

[134] G. Jacuzzi, S. Brazzola, and J. Kares, “Approaching immersive 3D audio broadcast streams of live performances,” *142nd Audio Engineering Society International Convention 2017, AES 2017*, 2017.

[135] K. Wise, “Creating Immersive Performances using Virtual Reality Technologies,” *University of Worcester Academic Blog*.

[136] B. Spiegel, G. Fuller, M. Lopez, T. Dupuy, B. Noah, A. Howard, and M. e. a. Albert, “Virtual reality for management of pain in hospitalized patients: A randomized comparative effectiveness trial,” *PLoS ONE*, vol. 14, no. 8, pp. 1–15, 2019.

[137] M. P. White, N. L. Yeo, P. Vassiljev, R. Lundstedt, M. Wallergå rd, M. Albin, and M. Lõhmus, “A prescription for “nature” - the potential of using virtual nature in therapeutics,” pp. 3001–3013, 2018.

[138] R. Alturki and V. Gay, “Augmented and virtual reality in mobile fitness applications: a survey,” in *Applications of Intelligent Technologies in Healthcare* (F. Khan, M. Jan, and M. Alam, eds.), Springer, 2019.

[139] T. Hilal Kadhim, S. Hekmat Salman, S. Khaled Khazaal, S. Salim Mohammed, A. Fadhil Hammad, and W. Saad Nsaif , “Survey the impact of the virtual reality in the fitness science and trainer’s performance,” *International Journal of Computer Science and Mobile Computing*, vol. 8, no. 1, pp. 1–7, 2019.

[140] T. Curtis, “Could realistic, advanced physics be the gameplay differentiator to take VR mainstream?,” *VR Focus*.

[141] D. Jagneaux, “Blood, Sweat, and Physics: How Boneworks turns your body into its key VR game mechanic,” *Upload VR*.

[142] J. Rees-Jones, J. Brereton, and D. Murphy, “Spatial audio quality and user preference of listening systems in video games,” in *18th Int. Conference on Digital Audio Effects (DAFx-15)*, pp. 223–230, 11 2015.

[143] J. Brereton, D. Murphy, and D. Howard, “The Virtual Singing Studio: a loudspeaker-based room acoustics simulation for real-time musical performance,” pp. 1–8, 2012.

[144] I. Neoran, M. Ben-Asher, and G. Alchanati, "Virtual reality music in the real world," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*, Aug 2020.

[145] B. Blesser and L. R. Salter, *Spaces Speak, Are You Listening? Experiencing Aural Architecture*. The MIT Press, 1993.

[146] M. Dobson, "Spaces Speak, Are You Listening? Experiencing Aural Architecture by Barry Blesser and Linda-Ruth Salter," *British Journal of Music Education*, vol. 25, no. 2, p. 207–209, 2008.

[147] Bela, "The Sound of Other Realities - Prototyping Spatial Audio for VR/AR with Bela Mini," 2018. Accessed: 05/06/2022.

[148] V. Hazan and R. Baker, "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2139–2152, 2011.

[149] D. Pelegrín-García, B. Smits, J. Brunskog, and C.-H. Jeong, "Vocal effort with changing talker-to-listener distance in different acoustic environments," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 1981–1990, 2011.

[150] I. 9921:2002(E), "Ergonomics–Assessment of speech communication," Standard, International Organization for Standardization, Geneva, CH, 2002.

[151] P. Kirby, *The Evolution and Decline of the Traditional Recording Studio*. PhD thesis, University of Liverpool, 2015.

[152] Cockos Incorporated, "REAPER: Digital Audio Workstation." reaper.fm.

[153] P. Virostek, “The Quick & Easy Way to Create Impulse Responses,” *Creative Field Recording*, 2014. Accessed: 11/2019.

[154] Wolfgang Amadeus Mozart, “K. 361 Mozart Serenade No. 10 in B-flat major, III Adagio,” *The Best of the Complete Mozart Collection Vol. 5: Serenades.* CD, Philips.

[155] J. Pätynen, B. F. Katz, and T. Lokki, “Investigations on the balloon as an impulse source,” *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. EL27–EL33, 2011.

[156] Impulse Record, “Convology XT Plugin.” https://impulserecord.com/project/convology-xt-plugin/.

[157] Facebook 360, “Facebook 360 Spatial Workstation.” https://facebookincubator.github.io/facebook-360-spatial-workstation/, 2020.

[158] Adobe, “Adobe Premiere Pro.” https://www.adobe.com/uk/products/premiere.html, 2020.

[159] VR New Technology HTC, “Vive cinema.” https://github.com/openbigdatagroup/vivecinema, 2019. Accessed: 14/12/2019.

[160] P. Boersma and V. van Heuven, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9-10, pp. 341–347, 2001.

[161] P. Boersma, “Acurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.,” *Institute of Phonetic Sciences*, 1993.

[162] Penguin Books, *The Penguin Dictionary of Physics.* 1991.

[163] Portland State University, "Single-Group Statistical Tests with a Binary Dependent Variable." https://web.pdx.edu/~newsomj/uvclass/ho_z-test.pdf, 2020.

[164] Unity, "Audio and video roadmap." https://unity.com/roadmap/unity-platform/audio-video, 2022. Accessed: 07/09/2022.

[165] Unity, "Ambisonic audio." https://docs.unity.cn/2020.2/Documentation/Manual/AmbisonicAudio.html, 2020. Accessed: 04/04/2020.

[166] B. LaBelle, *Background Noise: Perspectives on Sound Art.* Music and sound studies, Continuum International, 2006.

[167] Google, "Unity." https://resonance-audio.github.io/resonance-audio/develop/unity/getting-started.html, 2018. Accessed: 24/09/2020.

[168] A. Çamcı, "Some considerations on creativity support for VR audio," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1500–1502, 2019.

[169] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9701–9707, 2020.

[170] L. Pouru, "The parameters of realistic spatial audio: An experiment with directivity and immersion," 2019.

[171] M. Chion, *Audio Vision: Sound on Screen.* New York: Columbia University Press, 1994.

[172] M. Chion, *Film: A Sound Art.* Columbia: Columbia University Press, 2009.

[173] D. T. Murphy, S. B. Shelley, A. Foteinou, J. Brereton, and H. Daffern, "Acoustic heritage and audio creativity: the creative application of sound in the representation, understanding and experience of past environments," *Internet Archaeology*, June 2017.

[174] B. Katz, D. Murphy, and A. Farina, "Exploring cultural heritage through acoustic digital reconstructions," *Physics Today*, vol. 73, 12 2020.

[175] A. Bevilacqua, F. Merli, A. Farina, E. Armelloni, A. Farina, and L. Tronchin, "3dof representation of the acoustic measurements inside the comunale-pavarotti theatre of modena," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–4, 2021.

[176] A. Bevilacqua, F. Merli, L. Tronchin, and A. Farina, "Acoustic measurements of the Roman theatre of Pompei by mapping the sound reflections," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–5, 2021.

[177] L. Lavagna, L. Shtrepi, A. Farina, A. Bevilacqua, and A. Astolfi, "Virtual reality inside the greek-roman theatre of tyndaris: comparison between existing conditions and original architectural features.," in *PROCEEDINGS of the 2nd Symposium: The Acoustics of Ancient Theatres 2022*, 2022.

[178] J. Banerjee, "Royal Holloway, University of London," *The Victorian Web*, 2021.

[179] "The Choir of Royal Holloway," *The Choir of Royal Holloway*, 2022.

[180] A. Siddig, A. Ragano, H. Z. Jahromi, and A. Hines, "Fusion Confusion: Exploring Ambisonic Spatial Localisation for Audio-Visual Immersion Using the McGurk Effect," in *Association for Computing Machinery*, MMVE '19, (New York, NY, USA), p. 28–33, 2019.

[181] A. N. Nagele, V. Bauer, P. G. T. Healey, J. Reiss, H. Cooke, T. Cowlishaw, C. Baume, and C. Pike, "Interactive audio augmented reality in participatory performance," *Frontiers in Virtual Reality*, vol. 1, p. 46, 2021.

[182] A. Farina, "Aurora 2.4.1." http://pcfarina.eng.unipr.it/Aurora_XP/index.htm. Accessed: 28/06/2021.

[183] A. Gungormusler, "Resonance audio source code," *GitHub*, Mar 2018.

[184] L. Peng, "13 - sound absorption and insulation functional composites," in *Advanced High Strength Natural Fibre Composites in Construction* (M. Fan and F. Fu, eds.), pp. 333–373, Woodhead Publishing, 2017.

[185] A. J. Zuckerwar, "Acoustical measurement," in *Encyclopedia of Physical Science and Technology (Third Edition)* (R. A. Meyers, ed.), pp. 91–115, New York: Academic Press, third edition ed., 2003.

[186] Resonance Audio. https://resonance-audio.github.io/resonance-audio/, 2022.

[187] Xmaple, "Resonance Audio, spatialization trouble with too many audiosources," *Stack Overflow Questions*, 2019. Accessed: 17/08/2021.

[188] J. Cook and S. Mirashrafi, "Point cloud to sound cloud digital innovation and historic sound at linlithgow palace," *magazén*, dec 2022.

[189] Google, "Developer guide for resonance audio for unity," *GitHub repository*, 2018. Accessed: 21/03/2021.

[190] "Pure data," *Institute for Electronic Music and Acoustics*, 2004. Accessed: 1/10/2020.

[191] N. Moody, “LibPd Unity Integration.” https://github.com/LibPdIntegration/LibPdIntegration, 2021. Accessed: 12/07/2021.

[192] “Microphone - class in UnityEngine,” *Unity Documentation*, 2022. Accessed: 12/07/2021.

[193] M. Tippach, “ASIO4ALL Official Home,” *Steinberg Media Technologies GmbH*, 2022.

[194] M. Omer, L. Margetts, M. H. Mosleh, and L. S. Cunningham, “Inspection of concrete bridge structures: Case study comparing conventional techniques with a virtual reality approach,” *Journal of Bridge Engineering*, vol. 26, no. 10, p. 05021010, 2021.

[195] R. Kovask, “Survey of state of the art VR Driving Simulation for Physical Test Car Using LiDAR for Mapping the Surrounding Environment,” 2021.

[196] S. I. Zolanvari, D. F. Laefer, and A. S. Natanzi, “Three-dimensional building façade segmentation and opening area detection from point clouds,” *ISPRS journal of photogrammetry and remote sensing*, vol. 143, pp. 134–149, 2018.

[197] G. M. Akselrod, “Meta-Lidar: the sensor technology that makes AR/VR experiences real,” in *SPIE AR, VR, MR Industry Talks 2022*, vol. 11932, International Society for Optics and Photonics, SPIE, 2022.

[198] R. Tredinnick, G. Casper, C. Arnott-Smith, A. Peer, and K. Ponto, “Using virtual reality to study health in the home,” in *2018 IEEE Workshop on Augmented and Virtual Realities for Good (VAR4Good)*, pp. 1–5, 2018.

[199] B. Saylor, “pd-externals/bsaylor/readme.txt.” https://github.com/pd-externals/bsaylor/blob/master/README.txt, 2018. Accessed: 09/09/2021.

[200] B. Saylor, "help-partconv .pd." https://puredata.info/author/bensaylor, 2007.

[201] "All about Bela," Accessed: 23/03/2021.

[202] W. G. Gardner, "Efficient convolution without input/output delay," in *97th Audio Engineering Society Convention*, Audio Engineering Society, 1994.

[203] A. McPherson, "Phase vocoder part 1." https://learn.bela.io/tutorials/c-plus-plus-for-real-time-audio-programming/phase-vocoder-part-1/, 2020.

[204] A. McPherson, "Circular buffers." https://learn.bela.io/tutorials/c-plus-plus-for-real-time-audio-programming/circular-buffers/, 2020.

[205] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing.* Pearson Education Limited, 3rd ed., 2014.

[206] F. Roberts, "realtime-audio-convolver-bela," *GitHub*, 2023.

[207] P. Doyle, *Echo and reverb: fabricating space in popular music recording, 1900-1960.* Music/culture, Wesleyan University Press, 1st. ed., 2005.

[208] HTC and Valve, "HTC Vive VR Headset," 2016. accessed July 1st 2022.

[209] A. Politis, S. Tervo, and V. Pulkki, "Compass: Coding and multidirectional parameterization of ambisonic sound scenes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6802–6806, 2018.

[210] L. McCormack and A. Politis, "Spatial Audio Real-Time Applications (SPARTA)," *GitHub repository*, 2021. Accessed: 06/07/2022.

[211] M. Lester and J. Boley, "The effects of latency on live sound monitoring," *Journal of the Audio Engineering Society*, october 2007.

[212] Bela, “Live sound spatialisation: I - building an interface for multichannel sound diffusion,” 2022. Accessed: 05/06/2022.

[213] Bela, “The intimate earthquake archive - interactive wearable artwork where audiences experience seismic activity through their body,” 2022. Accessed: 01/09/2022.

[214] Audiolab, “SADIE: Spatial Audio for Domestic Interactive Entertainment,” 2017. Accessed: 10/01/2021.