The Impact of Multichannel Game Audio on the Quality of Player Experience and In-game Performance

Joseph David Rees-Jones

PhD

UNIVERSITY OF YORK

Electronic Engineering

July 2018

Abstract

Multichannel audio is a term used in reference to a collection of techniques designed to present sound to a listener from all directions. This can be done either over a collection of loudspeakers surrounding the listener, or over a pair of headphones by virtualising sound sources at specific positions. The most popular commercial example is surround-sound, a technique whereby sounds that make up an auditory scene are divided among a defined group of audio channels and played back over an array of loudspeakers. Interactive video games are well suited to this kind of audio presentation, due to the way in which in-game sounds react dynamically to player actions. Employing multichannel game audio gives the potential of immersive and enveloping soundscapes whilst also adding possible tactical advantages. However, it is unclear as to whether these factors actually impact a player's overall experience. There is a general consensus in the wider gaming community that surround-sound audio is beneficial for gameplay but there is very little academic work to back this up. It is therefore important to investigate empirically how players react to multichannel game audio, and hence the main motivation for this thesis. The aim was to find if a surround-sound system can outperform other systems with fewer audio channels (like mono and stereo). This was done by performing listening tests that assessed the perceived spatial sound quality and preferences towards some commonly used multichannel systems for game audio playback over both loudspeakers and headphones. There was also a focus on how multichannel audio might influence the success of a player in a game, based on their in-game score and their navigation within a virtual world. Results suggest that surround-sound game audio is preferable over more regularly used two-channel stereo systems, because it is perceived to have higher spatial sound quality and there is an improvement in player performance. This illustrates the potential for multichannel game audio as a tool to positively influence player experiences, a core goal many game designers strive to achieve.

Contents

Abstract **2**

Table of Contents **3**

List of Tables **9**

List of Figures **10**

Acknowledgements **13**

Declaration of Authorship **14**

1 Introduction **15**

1.1 Statement of Hypothesis 17

1.2 Description of Hypothesis 17

1.3 Statement of Ethics 18

1.4 Thesis Structure 18

1.5 Contributions to the Field 20

2 Concepts for Spatial Hearing **22**

2.1 The Transmission of Sound 22

2.2 The Human Hearing System 24

2.3 Sound Intensity and Sound Pressure 26

2.4 The Frequency Domain 28

2.5 Room Acoustics 29

2.6 Binaural hearing 33

2.6.1 Interaural Time Difference (ITD) 34

2.6.2 Interaural Level Difference (ILD) 36

2.6.3 Spectral Cues 37

2.6.4 Head-related Transfer Function (HRTF) . . . 38
2.7 Summary . . . 42

3 Multichannel Audio Playback **43**
3.1 Mono . . . 44
3.2 Stereo . . . 44
3.3 Surround-sound . . . 47
3.3.1 Surround-sound codecs . . . 50
3.3.2 Stereo Mix-down . . . 52
3.4 Virtual Home Theatre (VHT) . . . 53
3.5 Ambisonics . . . 54
3.6 Spatial Attributes . . . 55
3.6.1 Distance . . . 56
3.6.2 Localisation . . . 56
3.6.3 Depth . . . 57
3.6.4 Width . . . 58
3.6.5 Envelopment . . . 58
3.7 Summary . . . 59

4 Multichannel Video Game Audio **61**
4.1 A History of Multichannel Audio in Gaming . . . 61
4.2 Observations on Multichannel Game Audio . . . 67
4.2.1 Formats . . . 68
4.2.2 Perspective . . . 70
4.2.3 Use of Centre Channel . . . 73
4.2.4 Music Panning . . . 75
4.3 Video Game Content Case Studies . . . 76
4.3.1 Game Selection Criteria . . . 77
4.3.2 The Last of us: Remastered . . . 78
4.3.3 Alien: Isolation . . . 80
4.3.4 P.T. . . . 83
4.3.5 Ratchet and Clank: Tools of Destruction . . . 85
4.4 Summary . . . 87

5 Determining the Player Experience **89**

5.1 Quality of Experience . . . 90
5.1.1 QoE Definition . . . 90
5.1.2 Utilitarian QoE assessment . . . 91
5.1.3 Analytic QoE assessment . . . 93
5.1.4 QoE test framework . . . 94
5.2 Influencing Factors on QoE . . . 95
5.2.1 Audio as an influence on QoE . . . 97
5.2.2 Spatial audio as an influence on audio quality . . . 98
Loudspeaker Rendering . . . 100
Headphone Rendering . . . 101
5.2.3 Game Audio Quality . . . 102
5.3 Summary . . . 103

6 Perceived Spatial Quality and Player Preferences **104**

6.1 Research Question . . . 105
6.2 Pilot Study . . . 105
6.2.1 Method . . . 106
6.2.2 Pilot Study Design . . . 106
Participants . . . 106
6.2.3 Pilot Study Procedure . . . 107
Materials . . . 107
Questionnaire . . . 109
6.2.4 Pilot Study Analysis of Results . . . 111
Perceived Spatial Quality . . . 112
Preference . . . 112
6.2.5 Pilot Study Discussion . . . 114
6.3 Main Experiment . . . 115
6.3.1 Method . . . 115
6.3.2 Experimental Design . . . 115
Participants . . . 117
6.4 Experimental Procedure . . . 117
6.5 Results Analysis . . . 118

6.5.1 Overall spatial sound quality . . . 119
6.5.2 Individual attribute scores comparison . . . 119
6.5.3 Preference . . . 122
6.6 Discussion . . . 123
6.7 Summary . . . 126

7 Headphone Based Audio Rendering and Player Preferences 128
7.1 Research Question . . . 129
7.2 Method . . . 129
7.3 Experimental Design . . . 130
7.3.1 Participants . . . 130
7.4 Experimental Procedure . . . 131
7.4.1 Materials . . . 131
VHT Rendering Process . . . 132
Stereo Mix-down Process . . . 134
7.4.2 Questionnaire . . . 135
7.5 Analysis of Results . . . 135
7.6 Discussion . . . 137
7.7 Summary . . . 140

8 The Impact of Multichannel Audio on Player Performance 141
8.1 Research Question . . . 143
8.2 Method . . . 143
8.3 Experimental Design . . . 144
8.3.1 Participants . . . 145
8.4 Experimental Procedure . . . 145
8.4.1 Materials . . . 146
8.4.2 Loudspeaker Game Audio Rendering . . . 148
Sound Spatialisation . . . 150
Distance Attenuation . . . 150
8.4.3 Unity Headphone Plug-in . . . 151
Gathering of BRIR . . . 153
8.4.4 Training Session . . . 153

8.4.5 Apparatus . . . 154
8.5 Analysis of Results . . . 155
8.5.1 Player scores . . . 156
Group A . . . 156
Group B . . . 157
8.5.2 Route Directness Index . . . 158
8.5.3 Player Preference . . . 160
8.6 Discussion . . . 162
8.7 Summary . . . 167

9 Summary and Conclusions **169**
9.1 Chapter 6 - Perceived Spatial Quality and Player Preferences . . . 169
9.2 Chapter 7 - Headphone-Based Audio Rendering and Player Preferences . 170
9.3 Chapter 8 - The Impact of Multichannel Audio on Player Performance . . 171
9.4 Conclusions . . . 171
9.5 Further Work . . . 174
9.6 Final Remarks . . . 176

A Perceived Spatial Quality and Player Preferences Experiment Pack **178**
A.1 Experiment Information Sheet . . . 179
A.2 Consent Form . . . 180
A.3 Spatial Attributes Reference . . . 181
A.4 Event Time Line . . . 182
A.5 Control Scheme . . . 183
A.6 Questionnaire . . . 184
A.7 Demographic Information . . . 185
A.8 Participant comments . . . 186

B Headphone Based Audio Rendering and Player Preferences Experiment Pack **187**
B.1 Experiment Information Sheet and Consent Form . . . 188
B.2 Demographic Information . . . 189
B.3 Event Time Line . . . 190
B.4 Spatial Attribute Reference . . . 191
B.5 Questionnaire . . . 192

B.6 Participant Comments 193

C The Impact of Multichannel Audio on Player Performance Experiment Pack 194

C.1 Experiment Information Sheet 195

C.2 Consent Form 196

C.3 Questionnaire and Demographic Information 197

C.4 Participant Comments 198

C.4.1 Group A - Loudspeaker Playback 198

C.4.2 Group B - Headphone Playback 200

D Data CD Index 203

D.1 Chapter 6 Data and Analysis 203

D.2 Chapter 7 Max/MSP Patch 203

D.3 Chapter 7 Data and Analysis 203

D.4 Chapter 8 Game 203

D.5 Chapter 8 Data and Analysis 204

Glossary 205

Bibliography 208

List of Tables

2.1 Sound examples and associated SPL. 27

5.1 Audio attribute 'families'. 100

6.1 Pilot test spatial quality sign-test output. 112
6.2 Pilot test preference sign-test output. 113
6.3 Discrete channel routing to loudspeakers. 116
6.4 Participant group allocation. 118
6.5 Main experiment overall spatial quality sign-test output. 119
6.6 Individual spatial quality comparison between BMo and BSt. 121
6.7 Individual spatial quality comparison between BMo and 7.1. 122
6.8 Individual spatial quality comparison between BSt and 7.1. 122
6.9 Main experiment preference sign-test output. 123

7.1 Sign-test output for preference and spatial attribute scores. 136

8.1 Counterbalanced participant groups. 144
8.2 Loudspeaker player score sign-test output. 158
8.3 Headphone player score sign-test output. 159
8.4 Loudspeaker RDI sign-test output. 160
8.5 Headphone RDI sign-test output. 163
8.6 Loudspeaker preference percentages. 163
8.7 Headphone preference percentages. 164

9.1 Surround-sound definitions. 169

List of Figures

2.1 Compressions and rarefactions illustrated via golf balls and springs. . . . 23
2.2 The human hearing system. . . . 24
2.3 The uncoiled cochlea. . . . 25
2.4 The inverse square law related to sound waves. . . . 26
2.5 The relationship between the time and frequency domain. . . . 29
2.6 Sound wave reflections in a room. . . . 30
2.7 The relationship between the direct sound, early reflections and reverb. . . . 31
2.8 Simplified impulse response. . . . 32
2.9 Illustration of ITD. . . . 35
2.10 The cone of confusion 35
2.11 Head-shadowing. . . . 36
2.12 A sound wave interacting with the pinna. . . . 38
2.13 An example HRIR plot. . . . 39
2.14 An example HRTF plot. . . . 40
2.15 KEMAR mannequin. . . . 41

3.1 Coincident pair of figure 8 microphones. . . . 45
3.2 Phantom imaging. . . . 47
3.3 Surround-sound loudspeaker arrangements. . . . 49
3.4 Sampling an analogue signal. . . . 51
3.5 Illustration of audio depth. . . . 57
3.6 Illustration of audio width. . . . 59

4.1 Multichannel game audio timeline. . . . 62
4.2 Pong (video game). . . . 64
4.3 King Arthur's World (video game). . . . 65
4.4 Halo: Combat Evolved (video game). . . . 66

4.5 Doom 2016 (video game). 71
4.6 The Witcher 3: Wild Hunt (video game). 72
4.7 The Last of Us: Remastered (video game). 79
4.8 Alien: Isolation (video game). 81
4.9 Alien: Isolation radar system. 82
4.10 P.T. (video game). 84
4.11 Ratchet and Clank: Tools of Destruction (video game). 86

5.1 The MURAL model. 99
5.2 The IEZA framework. 103

6.1 First level of The Last of Us: Remastered. 107
6.2 Listening room. 108
6.3 Pilot test spatial quality boxplot. 113
6.4 Pilot test preference box-plot. 113
6.5 ITU 7.1 surround-sound loudspeaker angles. 117
6.6 Main experiment overall spatial quality boxplot 120
6.7 Individual attribute rating boxplot. 121
6.8 Main experiment preference boxplot. 123

7.1 Max/MSP block diagram. 133
7.2 Boxplots for preference and spatial attribute scores. 137

8.1 Top-down screenshot of custom game. 147
8.2 First-person screenshot of custom game. 147
8.3 Custom game concept. 148
8.4 Experiment loudspeaker angles. 149
8.5 Max/MSP and UDP signal flow. 149
8.6 Unity audio plug-in. 151
8.7 Unity volume roll-off. 152
8.8 Dummy head used for BRIR measurements. 154
8.9 Loudspeaker player score boxplot. 157
8.10 Headphone player score boxplot. 158
8.11 RDI example plots. 160
8.12 Loudspeaker RDI boxplot. 161

8.13 Headphone RDI boxplot. 162
8.14 Preference bar charts. 164

Acknowledgements

I would first like to say that this thesis would not have been made possible without Professor Damian Murphy, whose supervision throughout the PhD process has been invaluable. His overwhelming optimism, encouragement, and provision of wintery pints, has been the driving force towards the completion of this thesis, as well as my own personal growth. Reflecting on my time under Damian's supervision, I feel I have become an overall more confident person, and for this I am truly grateful. I also want to express my warmest thanks to Dr. Jude Brereton, who has acted as an excellent thesis advisor throughout my PhD and, most importantly, provided an endless supply of cake. Thanks also to Michael Kelly as my industry supervisor for providing support and to DTS, inc. (part of Xperi) for funding this work and therefore enabling my research.

Thanks must also go to all the members of the Audio Lab, for creating a truly special working environment that is somehow professional, enjoyable and ridiculous all at the same time. Special thanks to Frank, particularly for picking me up in my time of need.

Finally, I would like to extend my thanks to the family and friends who have put up with me over the past few years, especially Amelia, who has made this final push altogether more bearable.

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 sources are acknowledged as References.

In addition, I declare that parts of this research have been presented at conferences held during the course of the research degree. The related publications are as follows:

J. Rees-Jones and D. T. Murphy, "The Impact of Multichannel Game Audio on the Quality and Enjoyment of Player Experience", in *Emotion and Video Game Soundtracking: then, now and next*, pp. 143-163, Springer, Cham, Switzerland, 2018, ISBN 978-3-319-72271-9.

J. Rees-Jones and D. T. Murphy, "A Comparison of Player Performance in a Gamified Localisation Task Between Spatial Loudspeaker Systems", in *Proc. of the 20th Int. Conf. on Digital Audio Effects (DAFx17)*, pp 329-336, Edinburgh, UK, Sep. 5-9, 2017.

J. Rees-Jones and D. T. Murphy, "Spatial Quality and User Preference of Headphone Based Multichannel Audio Rendering Systems for Video Games: A Pilot Study", AES 142nd International Convention, Berlin, Germany, May 20-23, 2017, Convention Paper 9772.

J. Rees-Jones, J. S. Brereton and D. T. Murphy, "Spatial audio quality and user preference of listening systems in video games", in *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, pp. 223-230, Trondheim, Norway, Nov. 30- Dec. 3, 2015.

Chapter 1

Introduction

Audio has been an integral part of the video game playing experience ever since its inclusion in arcade machines in the 1970s. Sound effects and music played back during gameplay, and often in-sync with player interactions, are used as a tool to inform the player of their actions, progress the narrative of story-driven titles, and also elicit emotional responses [1–4]. A classic example of the importance of video game audio is in the original *Space Invaders*, playable on arcade machines in the late 1970s [5]. As the player progresses through the game, alien space ships creep ever closer to the bottom of the screen, and the tempo of the soundtrack increases. This simple, but effective, manipulation provides the player with an aural indication as to their in-game achievements, whilst also creating an increasingly tense atmosphere. Game audio has only continued to evolve over the years, with advancements in multichannel audio implementation being a key factor. In addition to this, the games industry has become worth more than music and film combined, being a key financial export for the United Kingdom [6, 7].

A key development for video game sound has been the increasing utilisation of multichannel audio systems, which is the use of more than one audio channel to give the impression that sound is emanating from multiple directions. Key examples of such listening systems include stereo and variations on surround-sound. Surround-sound is particularly interesting in a gaming context as spatialised sound cues can be used to fully envelop the player in the aural game world, creating immersive and dynamic soundscapes that contribute to more engaging gameplay experiences. At the time of writing, a high majority of video game content is able to output multichannel audio

conforming to home-theatre listening standards, such as 5.1 and 7.1 surround-sound. More recently Dolby Atmos [8] has been employed in a handful of big-budget titles, including *Star Wars: Battlefront* (2015) [9] and *Overwatch* (2016) [10].

In the wider, often non-academic, gaming community there is a belief that, in comparison to stereo, surround-sound is beneficial to gameplay and is therefore desirable [11–16]. A survey distributed among video game players by Goodwin [17] also gives reason to believe that video game players consider surround-sound to be an important factor in a game experience. There is, however, very little work formally investigating this idea in the academic literature, and hence the motivation for this thesis. Work by Letowski [18], Rumsey [19], Berg [20] and Dewhirst [21], among others, have considered the influence that surround-sound audio systems have on listening experiences for static and non-interactive multimedia content, such as music and film. In many of these cases, listening to audio over a surround-sound playback system is shown to enhance the listening experience, and thus the overall experience of a user. For this thesis, the same concepts are applied, but in the context of interactive video games, extending this prior work and providing a novel contribution to the field. It is the hope that relating multichannel audio playback to player experience could provide an important step in the advancement of audio technologies and sound design for video games, especially if there is found to be a positive influence in the overall experience of playing a game, which is an accomplishment that all game designers strive to achieve.

One of the main obstacles to overcome in this work will be in the development of novel listening tests using video game content as experimental stimuli. The author feels it is important for potential participants to feel as though they are engaged in playing a game, whilst also being under experimental conditions. This will require participants to play significant segments of video games, allowing them to become fully immersed in the experience. Care will need to be taken in the approaches towards designing assessment methodologies for measuring the perceived audio quality of the compared listening systems. This, therefore, provides a secondary motivation for this work, where it is hoped that the advantages and disadvantages of both traditional and novel test methodologies, in the context of gaming, will become clear to the reader.

1.1 Statement of Hypothesis

The overall hypothesis that is considered throughout, and so informs this thesis is as follows:

> *The implementation of surround-sound in an interactive video game environment, rendered either over loudspeakers or headphones, will have a positive impact on a player's gaming experience in comparison to stereo or mono.*

1.2 Description of Hypothesis

Implementation of surround-sound

Surround-sound audio systems expand on stereo by utilising a greater number of audio channels that are routed to an equivalent number of loudspeakers placed at specific points around the listening space. In this thesis, surround-sound is defined as an audio system that makes use of more than two discrete audio channels. The definition does not include systems that route two or fewer channels of audio to two or more loudspeakers surrounding a listener. The use of multiple audio channels output from different loudspeakers means that discrete sounds can be played from multiple locations around the listener to give the impression of movement, location and size. The application has the potential to make audio material altogether more immersive, enveloping and engaging, especially as the number of channels surrounding the listener increases beyond two.

Interactive video game environments

One of the core differences between video games and other screen media (such as film) is that the player is required to interact directly with the virtual environment presented to them. This means player input directly influences how and when sound is heard, unlike in film where the soundtrack is usually pre-rendered, essentially making the player an active participant in the creation of a unique audio experience.

Loudspeaker and headphone rendering for video games

Video game audio can be transmitted to the player over either an array of loudspeakers, or a pair of headphones covering their ears. For multichannel audio playback in video games, the rendering to different channels usually needs to be achieved in real-time due to the interactive nature of the gameplay. Loudspeaker systems will often provide more distinct spatial information, in that sound source positions can be represented physically in the listening space, whilst headphone systems are more convenient due to equipment costs and practicality. Throughout this thesis various multichannel formats are compared.

The player experience

The player experience relates to how the player will react to the presented gameplay. For this thesis, this experience is inferred based on preferences between different listening systems, the perceived sound quality of these systems, and the performance of the player. Performance refers to how successful a player is in a particular game, which can be measured through something such as a high-score, along with the ways in which they interact with the virtual environment. It is important to note that for headphone-based surround-sound playback, it is necessary to process the audio based on psychoacoustic theory and this can result in a varied experience between different listeners (see Section 2.6.4).

1.3 Statement of Ethics

The experiments presented in this thesis, and the management of corresponding data, were approved by the University of York Physical Sciences Ethics Committee, with reference numbers reesjones150319 for the work presented in Chapters 6 and 7, and Appendices A and B, and Rees-Jones090217 for Chapter 8 and Appendix C.

1.4 Thesis Structure

The remainder of this thesis is split into eight chapters, the first three relating to background in the area and the remainder presenting original experimental work and conclusions. The chapters are summarised as follows.

Chapter 2 introduces the fundamental concepts of hearing, specifically in relation to how humans hear with two ears i.e. binaurally. This includes the transmission of sound in air, the anatomy of the human ear, sound wave interactions in a room and the various systems used in binaural hearing including interaural time and level differences and spectral cues caused by the pinnae. The chapter finishes by introducing the concept of capturing head-related transfer functions for the synthesis of spatialised sound environments.

Chapter 3 explores the ways in which multichannel audio can be presented to a listener. This includes definitions of commonly used listening formats such as stereo and 7.1 surround-sound. Both loudspeaker- and headphone-based listening systems are presented and the benefits/shortcomings for both are considered. The chapter ends with a list of spatial audio attributes, with descriptions, commonly used to assess the sound quality of multichannel listening systems.

Chapter 4 considers the ways in which multichannel audio (specifically stereo and surround-sound) has been used in the video game industry, and the current state-of-the-art. This includes a history of audio implementation in video games, the author's own observations regarding surround-sound practices in gaming and a series of case studies focusing on four commercially available video games. These case studies are used to determine appropriate stimuli for later experimental work.

Chapter 5 introduces and defines the concept of *Quality of Experience* (QoE), a term often associated with a user's judgement of a piece of multimedia content. Examples are provided to illustrate how enhanced audio quality might influence the overall QoE of a user and how this might be used in the context of gaming. A framework for investigations into QoE is given, as well as terminology often associated with QoE measurement methods/metrics.

Chapter 6 describes the first of a series of listening tests exploring how the perceptual characteristics of multichannel game audio might influence the player experience. This is based on a user's subjective opinion of the spatial sound quality of three different loudspeaker based listening systems - mono, stereo and 7.1 surround-sound - using spatial attributes and descriptors. Preference scores are also used to infer the degree to which the experience of one game session might change in comparison to another.

Chapter 7 expands on the previous chapter by considering headphone-based equivalents of stereo and 7.1 surround-sound. The experiment uses the same methodology defined in the previous chapter, but for only two experimental conditions - a stereo down-mix of 7.0 listening material and a virtual home theatre (VHT) rendering of 7.0 surround-sound. In order to render the conditions for headphone listening, a custom system is designed in Max/MSP, utilising the surround-sound output from a commercially available video game.

Chapter 8 presents the final experiment in this thesis, investigating the impact multichannel rendering has on the performance of a video game player. The experimental conditions assessed are stereo, 7.1 surround-sound and an octagonal array, each played back over both loudspeakers and headphones. Player performance is based on how many sound sources can be found in a virtual game environment within a strict time limit. The virtual path taken by the player is also recorded to explore how the spatial cues offered by different listening systems might influence in-game navigation.

Chapter 9 summarises the background and experimental work given in the thesis, and gives conclusions based on the original hypothesis. The chapter concludes with ideas for further work.

1.5 Contributions to the Field

The purpose of this thesis is to explore how multichannel audio is used in the context of video games, of which there is minimal prior work in the literature. Below is a list of novel contritions to the field, based on the work presented throughout this thesis.

- **A review of multichannel audio in video games.** This extends on work presented by Kerins [2, 3], providing a history of multichannel audio in the context of video games and a review of the techniques used up to the point of writing this thesis (2018).
- **Development of listening test methodologies to determine multichannel audio quality and user experience using video games as experimental stimuli.** The experimental work presented throughout uses both commercially available and custom made video games as test stimuli to assess the perceptual qualities of

multichannel listening systems. This is important, as video games will often stimulate more senses that just the hearing system, by offering visual and tactile feedback, as well as interaction.

- **For the experiments presented in this thesis, multichannel systems with a higher channel count (like surround-sound) do not always out-perform lower channel systems.** The results presented throughout are perhaps not as clear-cut as one might predict, as in some cases surround-sound systems are not perceived to out-perform some of the more commonly used systems, such as stereo. This generally goes against pre-conceptions of multichannel audio and it may be the case that using interactive stimuli as test material had a more significant impact on results than expected. This therefore opens up some interesting opportunities for investigating listening test design in the context of video games, and how the methodologies presented throughout this thesis might be improved.

The findings introduced here are presented in detail throughout this thesis and also in the following publications:

J. Rees-Jones and D. T. Murphy, "The Impact of Multichannel Game Audio on the Quality and Enjoyment of Player Experience", in *Emotion and Video Game Soundtracking: then, now and next*, pp. 143-163, Springer, Cham, Switzerland, 2018, ISBN 978-3-319-72271-9.

J. Rees-Jones and D. T. Murphy, "A Comparison of Player Performance in a Gamified Localisation Task Between Spatial Loudspeaker Systems", in *Proc. of the 20th Int. Conf. on Digital Audio Effects (DAFx17)*, pp 329-336, Edinburgh, UK, Sep. 5-9, 2017.

J. Rees-Jones and D. T. Murphy, "Spatial Quality and User Preference of Headphone Based Multichannel Audio Rendering Systems for Video Games: A Pilot Study", AES 142nd International Convention, Berlin, Germany, May 20-23, 2017, Convention Paper 9772.

J. Rees-Jones, J. S. Brereton and D. T. Murphy, "Spatial audio quality and user preference of listening systems in video games", in *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, pp. 223-230, Trondheim, Norway, Nov. 30- Dec. 3, 2015.

Chapter 2

Concepts for Spatial Hearing

Before investigating the ways in which audio may influence video game player experiences, it is first important to consider the way in which sound is transmitted, and the biological mechanisms in place that allow it to be perceived by a listener. The environment in which a sound interacts and the anatomy of a listener both play major roles in the fundamental perception of sound, especially in relation to hearing spatial characteristics such as directionality, distance and reverberation. This chapter will introduce some of the core concepts underpinning the transmission of sound as acoustic pressure waves and how these are decoded by the human hearing system. There is also a focus on how the two ears work to allow a listener to locate and localise sounds from around their environment. The timing, amplitude and spectral cues used for directional hearing are introduced, as well as the way in which these cues can be replicated over a pair headphones to synthesise spatialiased audio environments.

2.1 The Transmission of Sound

Sound is heard when a series of vibrations emitting from some object (i.e. the sound source) are transmitted through a medium, usually air, and processed by the brain, after being received at the ear. The vibrations from the source cause variations in pressure between subsequent molecules of the transmission medium as they are compressed together and pulled apart (rarefacted). For this reason sound is a type of longitudinal wave which is a type of wave that travels in the direction of propagation. As illustrated in Figure 2.1, when these changes in pressure are regular, or periodic, then simple waveforms such as sinusoids are generated. Howard and Angus [22] provide an

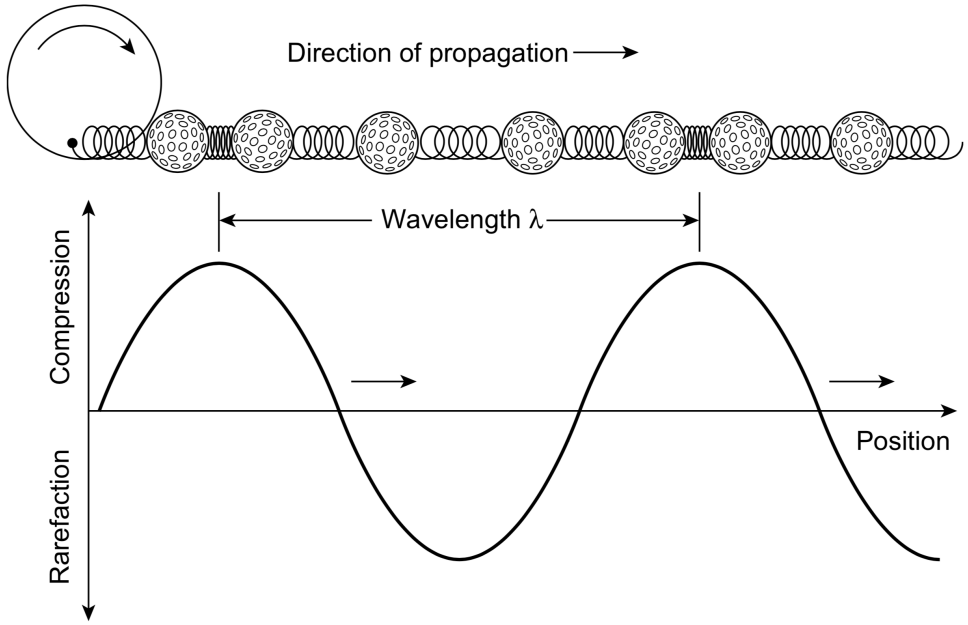


FIGURE 2.1: An illustration demonstrating the relationship between a longitudinal wave (the golf balls and springs) and transverse wave travelling in the direction of propagation, taken from [22].

analogy to illustrate this idea using golf balls as the molecules in the medium, and springs as the intermolecular forces between them. As the first golf ball is pushed to the right, this causes the spring next to it to compress and therefore pressure is increased. As the golf ball moves to the left, the spring rarefacts resulting in a decrease in pressure. As this process repeats, subsequent increases and decreases in pressure are generated in the the direction of propagation, which in this analogy is to the right. Figure 2.1 demonstrates how the compressions and rarefactions between air molecules (the y-axis) can be represented as a transverse wave travelling in the direction of propagation (the x-axis).

The period of a waveform is the time it takes for one complete repetition of the waveform. The number of periods to happen in one second defines the frequency of the waveform, measured in hertz (Hz). If the frequency (f) and the speed of sound in the medium (c) are known then it is possible to calculate the wavelength (λ), which is the distance between two successive peaks in the generated waveform. The importance of the wavelength in the context of binaural hearing will be discussed in Section 2.6, and is calculated using the following equation:

$$\lambda = \frac{c}{f} \tag{2.1}$$

If the transmission medium is air, which is the case in the majority of listening environments, then c will be equal to the speed of sound in air, approximately 344m/s for dry air at room temperature. Subsequently, if the wavelength is already known, then the equation can be rearranged to determine the frequency at that point in time:

$$f = \frac{c}{\lambda} \tag{2.2}$$

2.2 The Human Hearing System

The human hearing system is made up of three parts: the outer, middle and inner ear (see Figure 2.2), and sound is heard when the pressure waves transmitted from a sound source interact with it [23]. The outer ear comprises of the pinna and ear canal, with the collective purpose of funnelling incoming sound pressure waves to the remaining parts of the hearing system. The importance of the pinna, the protruding fold of flesh located

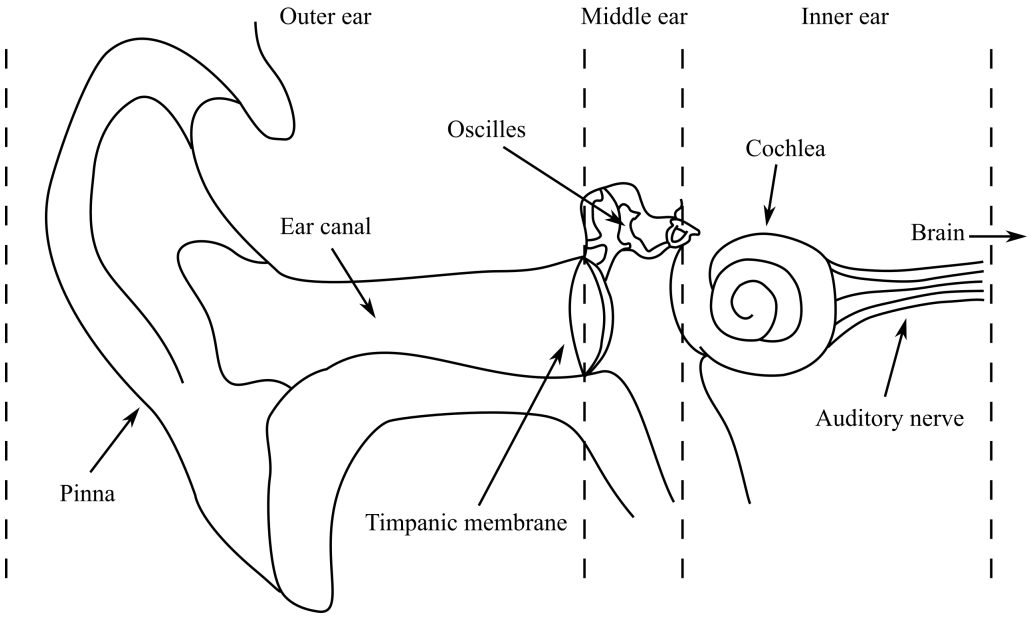


FIGURE 2.2: A simplified illustration of the human hearing system consisting of the outer, middle and inner ear adapted from [22]. Pressure waves are funnelled to the middle ear through the ear canal and subsequently converted into neural firings at the inner ear, which are transmitted to the brain via the auditory nerve. Once processed by the brain, this is perceived as sound.

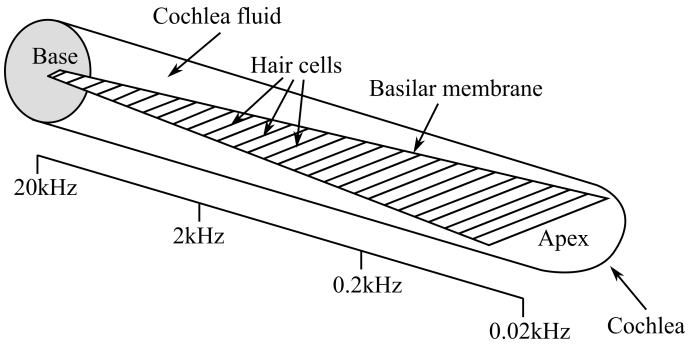


FIGURE 2.3: An illustration of the cochlea as it would look uncoiled with the basilar membrane running along its length with perpendicular lines representing sensitive hair cells. Note that the spacing between these lines is to illustrate the approximate placement of the hairs and is not to scale.

at the sides of the head, in relation to binaural hearing will be introduced in Section 2.6. The ear canal is a tube-shaped structure connecting the pinna to the timpanic membrane (or eardrum) which vibrates when excited by an incoming pressure wave. This causes the oscilles (three small bones located in the middle ear) to be pushed and pulled in a lever-like motion, serving to increase the amplitude of the pressure wave before being received at the inner ear [24].

The cochlea is a coiled structure located in the inner ear that converts the mechanical vibrations from the ossicles into neural firings that can be transmitted to the brain and perceived as sound. The incoming pressure waves are transmitted through a fluid within the cochlea with a higher impedance than air. The ossicles increase the amplitude of the incoming pressure wave before it reaches the cochlea such that it can be transmitted through this high impedance fluid, hence their importance [22]. The basilar membrane runs through the cochlea, which has a narrow base at the end closest to the ossicles and a wider apex at the other. To illustrate this, a diagram of an uncoiled cochlea is given in Figure 2.3. A distribution of sensitive hair cells also run along the length of the basilar membrane, collectively known as the organ of corti. As the amplified pressure wave is transmitted from the middle ear to the cochlea, the fluid contained within is displaced causing the basilar membrane to vibrate. When the frequency of the pressure wave is low, the peak of the vibration across the basilar membrane occurs towards the apex, whilst for higher frequencies this peak is located closer to the base [25]. The motion of the basilar membrane causes the hair cells along its length to flex, which is picked up by the auditory nerve and subsequently transmitted to the brain as neural firings. This is one of the core

mechanisms that allows for the perception of pitch, which relates to the arrangement of notes from low to high on a musical scale [26]. This is illustrated in Figure 2.3 where it can be seen that a peak in the waveform closer to the base will be perceived to be a lower pitch, whereas a peak located closer the the apex will be perceived to be a higher pitch.

2.3 Sound Intensity and Sound Pressure

The vertical axis in Figure 2.1 represents the amplitude of a waveform, which relates to the amount of force compressing and pulling apart the molecules of the transmission medium. For audio, the amplitude can be quantified as the sound intensity level (SIL), which corresponds to the amount of power, in watts, transmitted from the source per unit area (W/m^2) [22]. SIL relates to the perceived loudness of a sound, in that the higher the SIL the louder a sound will be to a listener. Due to the response of the ear being dynamic and non-linear, there is a large possible range in values for SIL, therefore it is often presented logarithmically in decibels (dB) [27]. SIL is defined as:

$$dB_{SIL} = 10log_{10}(\frac{I}{I_{ref}}) \tag{2.3}$$

where dB_{SIL} is the sound intensity in dB, I is the measured power density in W/m^2 and I_{ref} is a reference power density of $10^{-12}W/m^2$.

As a sound is transmitted from the source to a receiver it loses energy as a result of it spreading out in three-dimensions, as illustrated in Figure 2.4. This causes the intensity

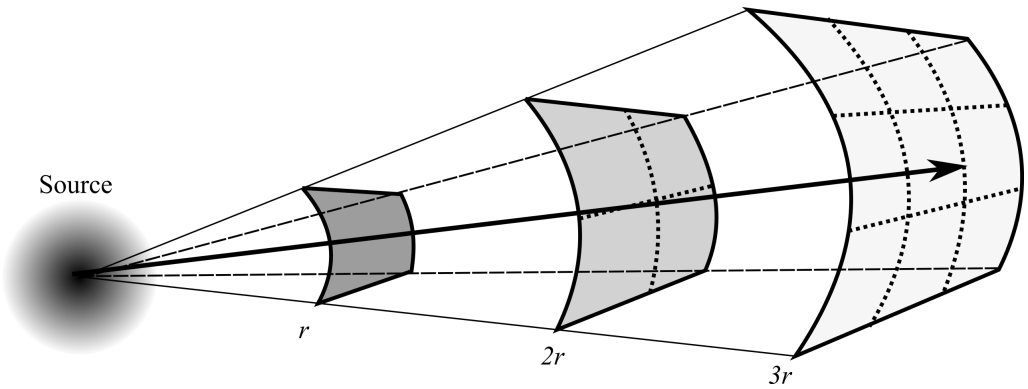


FIGURE 2.4: Illustration of sound intensity dropping in proportion to the amount of distance travelled (r). The lighter shades of grey illustrate the intensity level dropping as energy is spread outwards.

of the generated sound wave to be decreased, as the distance between the source and receiver increases and vice versa. This relationship conforms to the inverse square law, where the decrease in sound intensity is inversely proportional to the square of the distance (r) from the source [28]. The intensity of a sound radiating in three-dimensions, at distance r, is given by:

$$I = \frac{W}{4\pi r^2} \tag{2.4}$$

where I is the intensity level in W/m^2 and W is the power measured at the source in watts [22]. Based on this, for each doubling in distance, the intensity will reduce by a factor of 4, which is equivalent to $6dB_{SIL}$.

The sound pressure level (SPL) relates to the root mean square (rms) pressure of a waveform, at a specific point in time [22]. This quantifies the changes in pressure, measured in pascals (Pa), as a result of molecules in the transmission medium being compressed and rarefacted. Like SIL, SPL is given logarithmically in dB:

$$dB_{SPL} = 20log_{10}(\frac{P_{rms}}{P_{ref}}) \tag{2.5}$$

where dB_{SPL} is the sound pressure level in dB, P_{rms} is the root mean square of the pressure and P_{ref} is a reference equal to the pressure generated by a 1kHz tone at the threshold of human hearing (20μPa) [27]. Table 2.1 provides some reference SPL measurements to illustrate the relationship with perceived loudness.

Source/observing situation	Typical SPL in dB SPL
Hearing threshold	0dB
Whisper in quiet library at 6 feet	30dB
Normal conversation at 3 feet	60dB
City traffic (inside car)	85 dB
Jackhammer at 50 feet	95dB
Jet engine at 100 feet	140dB
Pain threshold	120-140dB

TABLE 2.1: A table of sound examples and the associated sound pressure level (SPL) in dB adapted from [29].

2.4 The Frequency Domain

In accordance to Fourier's theorem, complex waveforms are made up of multiple sines and cosines at varying frequencies, see Figure 2.5b. A waveform represented in the time domain (see Section 2.1) has an equivalent representation in the frequency domain which shows the frequency components (i.e. the constituent sines and cosines) in Hz along the x-axis and the amplitude of those individual components along the y-axis. A plot given in the frequency domain is usually referred to as a spectral plot, an example of which can be seen in Figure 2.5a for a single sine wave, oscillating at a constant frequency. For audio, the balance of frequency components contributes to the perception of timbre when paired with other time-varying factors such as the onset and offset of a waveform. The timbre of a sound is the specific tonal quality that allows it to be differentiated from others, independent of the pitch and loudness [31]. By performing a Fourier Transform on a complex waveform, it is decomposed into basis functions through correlation, i.e. the constituent sines and cosines, each with a frequency and amplitude [32]. The Fourier Transform of a waveform sampled at x is defined in [33] by:

$$X(\omega_k) = \sum_{n=0}^{N-1} x(t_n) e^{-j\omega_k t_n}, \qquad k = 0, 1, 2, ..., N-1 \tag{2.6}$$

where $X(\omega_k)$ is the spectrum of x, at frequency ω_k and $x(t_n)$ is the amplitude of x at time t_n in seconds. t_n is the nth sampling instant in seconds and ω_k is the kth frequency sample. N is the number of time samples in the input which is equal to the number of frequency samples in the spectrum at the output. This process can be reversed using an Inverse Discrete Fourier Transform (IDFT) whereby information in the frequency domain is transformed to the time domain, defined by:

$$x(t_n) = \frac{1}{N} \sum_{k=0}^{N-1} X(\omega_k) e^{j\omega_k t_n}, \qquad n = 0, 1, 2, ..., N-1 \tag{2.7}$$

where N is equal to the number of frequency samples [33].

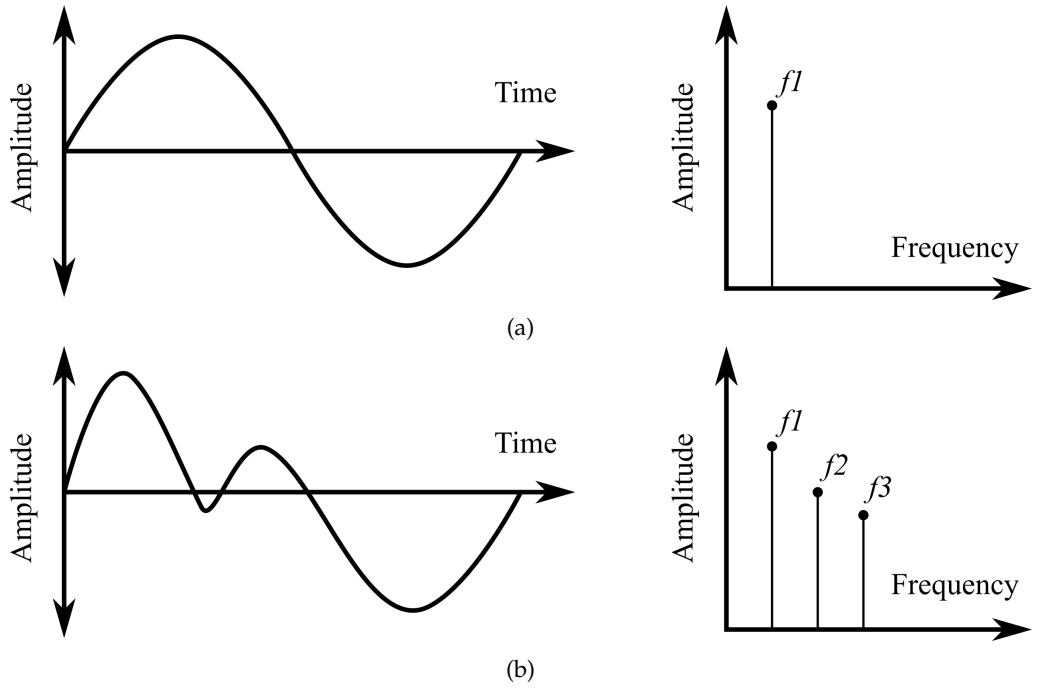


FIGURE 2.5: An illustration of the relationship between the time domain (left) and frequency domain visualised using a spectral plot (right), adapted from [30]. For the first plot (a) the waveform is a sinusoid, therefore only one frequency (*f1*) is given on the spectral plot. The second plot (b) is a more complex waveform, made up of multiple frequencies (*f1*, *f2* and *f3*) added together.

2.5 Room Acoustics

Up to this point, sound propagation has only been considered in the free field, meaning the transmission from source to receiver is not obstructed by the surrounding environment [30]. In real listening environments, it is likely that a sound pressure wave will interact with boundaries, such as walls, causing it to be reflected and energy to be absorbed. This alters the perceptual characteristics of the sound, such as the loudness and timbre, giving an impression of the shape and space in which the sound wave is propagating, as well as the distance between the source and listener. This is defined by three components: the direct sound, early reflections and reverberation (or late reflections), see Figure 2.6.

Direct sound refers to the shortest path between the source and receiver, where the sound wave is not affected by boundaries and obstructions. As introduced in Section 2.3, each doubling of distance will result in the intensity of a sound being reduced by 6dB. The distance also relates directly to the time it takes for a sound wave to reach the receiver, resulting in a delay between when the sound is emitted and when it is heard. This is

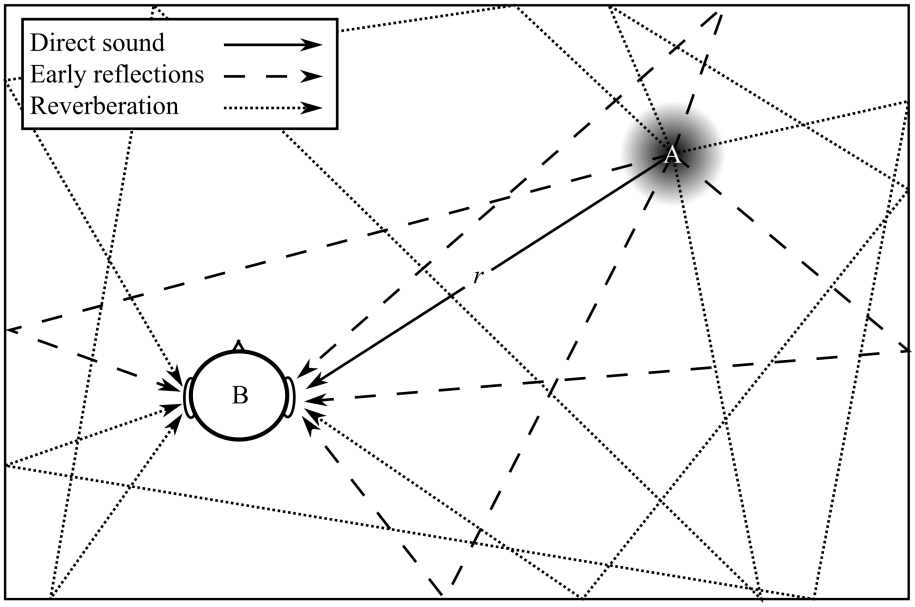


FIGURE 2.6: A top down illustration of the direct sound, early reflections and reverberation propagating through a room from a source (A) to a listener (B). The direct sound is unaffected by the boundaries of the room, whereas the early and late reflections occur when the energy of the sound wave is absorbed and subsequently reflected by these boundaries.

defined as:

$$\Delta t = \frac{r_d}{c} \tag{2.8}$$

where Δt is the time delay in milliseconds, r_d is the shortest distance between the source and receiver in meters and c is the speed of sound. The direct sound is an important indicator as to the perceived direction of a sound source relative to a listener due to the spacing between the two ears, and this will be discussed further in Section 2.6.

Reflections occur when a sound wave is incident upon a boundary, such as a wall, and changes direction as a consequence. When this happens, the amplitude of the reflected sound wave is less than that of the direct sound as energy is lost through being transferred and transmitted across the medium of the boundary [34]. The amount of absorption depends on the type of material the sound wave interacts with. For example, a hard surface, like a stone wall, will absorb less acoustic energy than a softer material, like a fabric curtain. As the number of subsequent reflections, and the total distance travelled, increases, the amplitude of the sound wave continues to drop until it is no

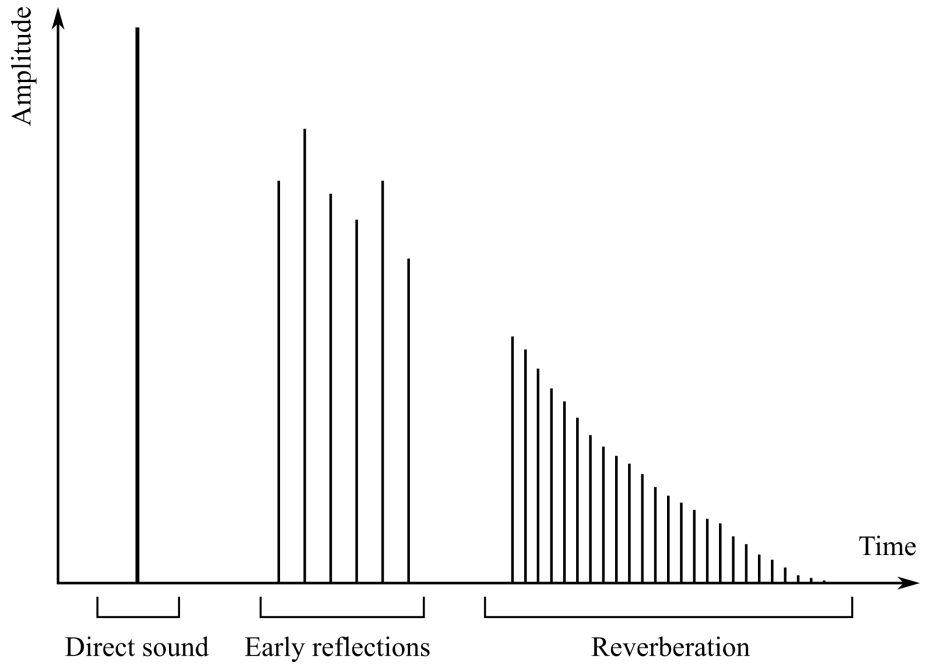


FIGURE 2.7: A simplified illustration showing the relationship between the amplitudes of direct sound, early reflections and reverberation. The direct sound has the highest amplitude because it represents the shortest path between the source and receiver. As the sound reflects on surfaces in the environment, energy is lost due to absorption from both air and the interacting materials, therefore the amplitude of each successive reflection reduces over time.

longer audible i.e. it is below the threshold of human hearing. The drop in amplitude of reflections over time is illustrated in Figure 2.7.

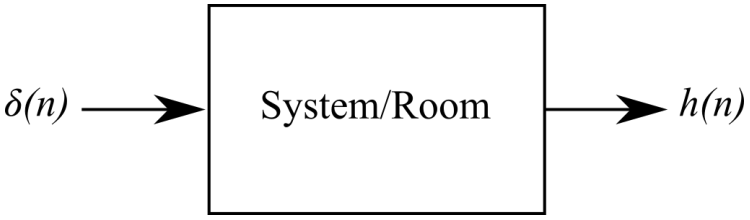
Early reflections are usually defined as those which arrive at the listener within 80ms of the direct sound; however, this is a somewhat arbitrary measure since the interaction with the boundaries of, and objects within, the surrounding environment are heavily influenced by the size and shape of the room. Recently researchers have proposed statistical measures to determine the boundary point between early reflections and late reverberation [35]. These reflections will have lost the least amount of energy, but are still generally lower in amplitude than the direct sound. The sound intensity of a reflection is defined as [22]:

$$dB_{SIL} = I_{source} - 20log_{10}(r_{ref}) + 10log(1 - \alpha) \tag{2.9}$$

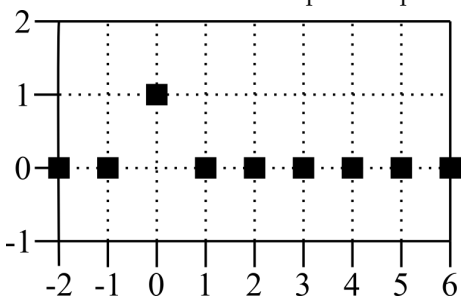
where dB_{SIL} is the sound intensity of the reflection in decibels, I_{source} is the measured sound intensity level of the source at 1 meter, r_{ref} is the total distance travelled and α is

an absorption coefficient between 0 and 1. Some common absorption coefficients for different materials are given in [36]. For example, a smooth marble floor has an absorption coefficient of 0.01 at 1000Hz whereas 9mm thick tufted carpet on a felt underlay has a coefficient of 0.6 at 1000Hz [37]. The reverberated energy will usually arrive at the ears after the early reflections (i.e. after at least 30ms) and will be a fusion of more closely packed reflections.

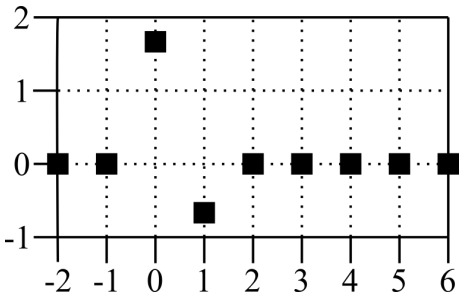
When the time delay between early reflections and the direct sound being received at a listener's ears is relatively short (less than 50ms [38]) they are perceived as one 'fused' sound. Even if the reflections arrive at the listener from multiple directions, this has no impact on the direction in which the original sound source is perceived to be, because the direct sound will be received at the ears first and is therefore most dominant. This is known as the Haas, or precedence, effect [39]. Longer delays result in perceptually distinct echoes, where the early reflections are heard separately after the direct sound, usually at a lower amplitude. Reverberation is the cluster of later reflections, and the time in which it takes for these reflections to become perceptually inaudible is the reverberation time [40]. Along with the amplitude cues introduced previously, the ratio in level between the direct sound and the reverberation is used as a cue for distance perception [41]. As the distance of a sound source relative to a listener increases, the



(a) An impulse ($\delta(n)$) excites a system which subsequently outputs the impulse response ($h(n)$).



(b) The delta function ($\delta(n)$).



(c) The impulse response ($h(n)$).

FIGURE 2.8: An illustration of a delta function (b) being used as the input to a system (a) to capture the impulse response at the output (c). Adapted from [32].

perceived loudness level drops due to energy being lost by the inverse square law as introduced previously. In reverberant spaces, the energy of the reverberance remains constant regardless of the distance between source and listener, resulting in a decrease in the direct-to-reverberant (D/R) energy ratio [42]. The perceived distance of a sound source will therefore increase as the D/R ratio decreases [43].

The acoustic qualities of a room can be captured by measuring the impulse response (IR). An IR is the signal that exits a system (i.e. a room) after it has been excited by an impulse, ideally a Dirac delta function which is defined as a single sample with a unit amplitude such that it contains equal energy at all frequencies [32]. The relationship between an impulse ($\delta(n)$) and the impulse response ($h(n)$) is presented in Figure 2.8. A room can also be excited using the swept sine method, whereby a a sine tone that increases in frequency is output from a loudspeaker and recorded using a microphone [44]. The range of frequencies is usually between 20Hz and 20kHz to account for the range of human hearing [40]. Once the sine sweep has been recorded, the IR is generated by convolving the recorded output signal with a version of the input signal that has been reversed in time [45]. The IR can then be convolved with a recorded audio sample to give the illusion that it is emanating from the space in which the IR was recorded.

2.6 Binaural hearing

Binaural hearing refers to how both the left and right ear work together in allowing a listener to perceive sound emanating from the space around them i.e. spatially. A listener is able to infer the direction of a sound in a procedure whereby the brain detects, and subsequently processes, small differences in the physical properties of a sound pressure wave between the two ears. The ear closest to a sound source is known as the ipsilateral ear, whilst the furthest is the contralateral. Differences between the two ears include properties such as time of arrival, amplitude and frequency content. This section introduces some of the core concepts of binaural hearing, and how the characteristics can be captured for a listener as their head-related transfer function (HRTF).

2.6.1 Interaural Time Difference (ITD)

For sounds that do not emanate from directly in-front of or directly behind the listener, a sound pressure wave will arrive at one ear before the other. This results in a small difference in the time of arrival between the two ears, know as the interaural time difference (ITD) [46]. ITD is one of the properties used by the brain to infer the direction of a sound source, with the pressure wave arriving first at the ear closest to the sound source (see Figure 2.9). Using Woodworth's formula [47], the ITD can be defined as:

$$ITD = d(\theta + sin\theta)/c \tag{2.10}$$

where the ITD is given in s (with values typically of the order of μs), d is half the distance between the ears in metres and θ is the angle of incidence in radians of the sound pressure wave relative to a forward-facing listener [22]. The maximum ITD (approximately 660μs) will occur when a sound source is positioned directly to the left or right of a listener, as a result of the pressure wave travelling the longest possible distance to reach the contralateral ear [38]. However, although the ITD at this position is relatively large, localisation errors can still occur due to the 'cone of confusion' [48], see Figure 2.10. The 'cone of confusion' is defined as the area around the ear in which incident sound waves will have the same ITD cues, making precise localisation of lateral and elevated sound sources difficult [49]. In addition to this, when the angle of incidence is 0°, or 180° (i.e. the sound is located directly in-front of or behind the listener's head) then the sound will be received at both ears simultaneously. In these cases, a listener may experience front/back confusion where it is difficult to differentiate a sound emanating from in-front or behind. By performing small head movements, it is possible to correct localisation errors caused within the 'cone of confusion' and by front/back confusion, in that the angle of incidence will change to one with a more useful ITD [38]. Spectral cues occurring as a result of a sound wave interacting with the pinnae also aid in the differentiation between front and rear sounds and are discussed in Section 2.6.3.

ITD is most useful when the wavelength (see Section 2.1) of the waveform is larger than the distance between the two ears, in other words when the frequency is low. This is due to the hearing system being sensitive to phase differences between the signals at

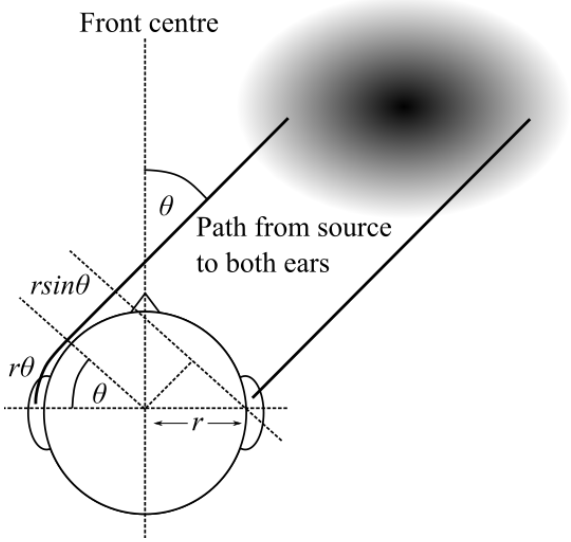


FIGURE 2.9: Illustration of a sound pressure wave emitted from a source arriving at one ear (right) before the other (left), resulting in a delay in the time of arrival between the two ears, i.e. the interaural time difference (ITD). Adapted from [46].

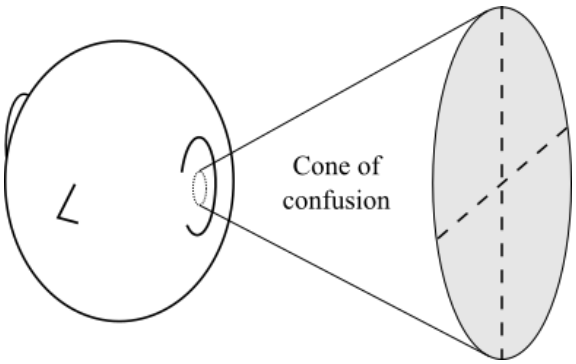


FIGURE 2.10: An illustration of the cone of confusion, where sounds that arrive at the ear within it will have the same ITD. Adapted from [38].

each ear, caused by the different path lengths between the sound source and the left and right ears. For frequencies above approximately 1500Hz, the wavelength is generally shorter than the diameter of the human head, which causes a greater phase shift of the waveform when it is received at the contralateral ear [38, 50]. Localisation then becomes more difficult because the separation between the onset of the waveform that arrives first and the one preceding it becomes more ambiguous. For this reason, other hearing mechanisms are believed to account for localisation at higher frequencies, such as level and spectral differences.

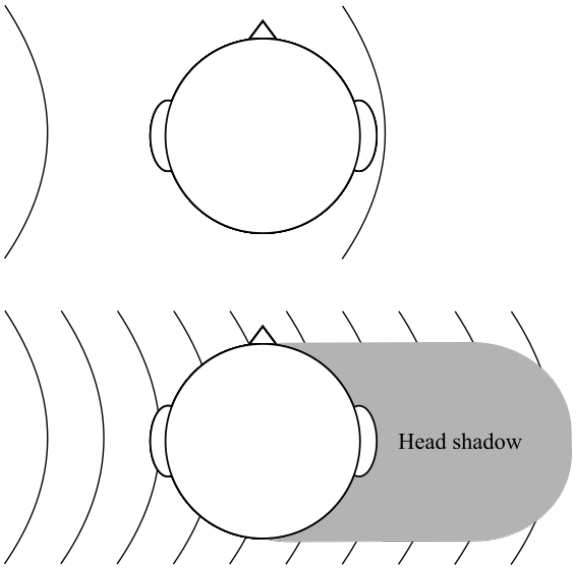


FIGURE 2.11: When the wavelength is larger than the diameter of the head, a waveform is diffracted to the contralateral ear. Higher-frequencies are attenuated by a head shadow effect, as illustrated in this diagram.

2.6.2 Interaural Level Difference (ILD)

For frequencies above about 1500Hz, the difference in the amplitude of a waveform between the two ears aids in the perception of directionality. This interaural level difference (ILD) is caused by the head acting as an absorbing barrier between the two ears, reducing the amount of energy in a sound pressure wave as it transmits to the contralateral ear [51]. This is known as a head-shadowing effect and is illustrated in Figure 2.11. This results in a sound being louder on one side of the head than the other, and thus the source is perceived to emanate from the direction in which the strongest signal is detected [52].

Lower frequencies with a wavelength larger than the diameter of the head will diffract around the head and therefore no head-shadowing will occur, meaning ILD is generally only useful for localising high frequencies [38]. This is the basis for the Duplex Theory proposed by Lord Rayleigh [53], which suggests that there is coordination between interaural cues in the localisation of sound sources, where the brain uses time differences to localise lower frequencies and level differences for higher frequencies.

2.6.3 Spectral Cues

The spectral content of a sound is also changed as it is transmitted from the source to the inner ear, and is another cue used by listeners to infer the location of a sound source, especially those located above or below the listener. Sound waves interact with anatomical features, such as the pinnae, torso and shoulders, causing them to reflect and diffract before being received at the timpanic membrane. Short delays occur relative to the direct sound, resulting in comb-filtering, meaning that equally spaced notches appear in the spectrum of a sound as a consequence of the delayed signal being out of phase with the original [54]. The spectral changes caused by the comb-filtering are decoded by the brain as directional information [55]. In the frequency domain, these changes are characterised as peaks (amplified frequencies), and notches (attenuated frequencies), relative to the unaltered pressure wave. Specific peaks and notches in the spectrum are determined by observing a listener's measured HRTF, discussed in Section 2.6.4.

The grooves and folds in the pinnae are believed to contribute significantly to the formation of spectral cues. This is because an incoming sound wave will interact with the shape of the pinnae differently, depending on the lateral and vertical position of the sound source, relative to a forward-facing listener [56]. Reflections from different points on the pinna will result in slightly different peaks and notches in the frequency domain, providing information regarding the direction from which the sound is emitting. The pinna also acts as a barrier for sounds located behind a listener, again altering the spectral content. An example of the difference in reflection between two sound source positions is given in Figure 2.12. Because elevated sound sources will provide similar interaural differences, it is the spectral changes caused by the pinnae that are used by a listener to infer the height of a sound source. In general, only frequencies above approximately 3kHz will interact with the pinnae due to the wavelength being comparable, or shorter than, the relatively small dimensions of the outer ear for most listeners [57]. The spectral changes are also dependant on the listener themselves, because no two pinnae, between different individuals, are identical. For this reason it is difficult to generalise spectral cues caused by the pinnae, because the peaks and notches in the spectrum will be unique for each listener.

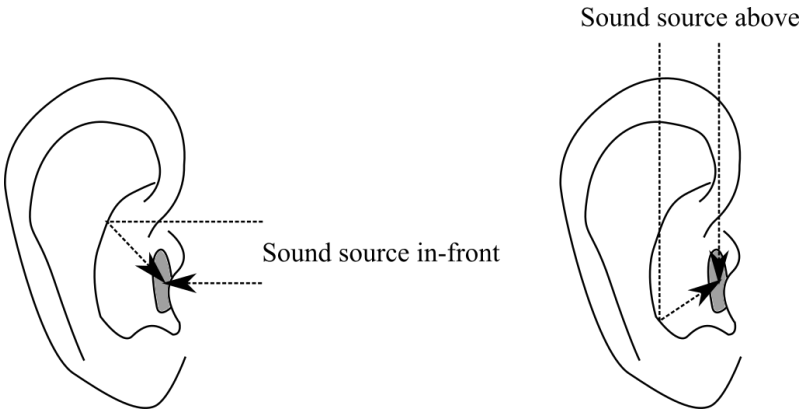


FIGURE 2.12: An illustration of how a sound positioned in-front of a listener might interact with a pinna (left image) compared to one positioned above the listener (right image), adapted from [57]. The reflections at different points on the pinnae form spectral cues used to infer the direction from which a sound is emitting.

2.6.4 Head-related Transfer Function (HRTF)

The interaural and spectral cues discussed previously combine to form a head-related transfer function (HRTF) for every possible sound source position around a listener. Since different sound source positions provide distinct interaural and spectral cues, the HRTF for each position will be unique and can be processed by the brain to infer directionality both laterally and vertically [46]. A measured HRTF is where the spectral peaks and notches caused by pressure waves interacting with the pinnae can be observed. As stated previously, the spectral changes from the pinnae cannot be easily generalised between individuals due to differences in the dimensions of the outer ear. For this reason, a consistent sound source position will result in a dissimilar HRTF between two listeners.

HRTF measurements are generated by converting a head-related impulse response (HRIR) from the time domain to the frequency domain via a Fourier transform (see Section 2.4). A HRIR pair (one for each ear) is obtained by recording a signal that is received at two small microphones placed inside, or near, a listener's left and right ear canal, effectively mimicking the function of the timpanic membrane [59]. The swept sine method introduced in Section 2.5 is appropriate, where the system under consideration is a listener's combined hearing system, rather than the environment in which a sound interacts [44, 60]. A set of unique HRIRs can be produced by outputting a signal, for example from a loudspeaker, from multiple positions around a listener. For each pair of HRIRs, the recorded signal will contain encoded directional information corresponding

to the azimuth and elevation from which the original signal was output. It is also recommended that a HRIR should be taken in a controlled and anechoic (non-reflective) environment, to ensure that potential reflections from surroundings do not colour the response [56]. Figure 2.13 is an example of a HRIR plotted in the time domain for the left (blue line) and right (orange line) ears for a sound source emanating to the left of the listener. This illustrates both the interaural time and level differences for a sound source at that position, where it can be seen that the onset of the left signal is before that of the right. The amplitude for the left ear signal is also higher than that of the right due to the head shadowing that occurs at the contralateral ear.

Figure 2.14 is an example of a pair of HRTFs derived by performing a Fourier transform on the same set of HRIRs presented in Figure 2.13. The figure illustrates the spectral notches that occur as a result of the sound wave interacting with the left pinna before being received at the right, which is overall lower in amplitude due to the head shadowing effect.

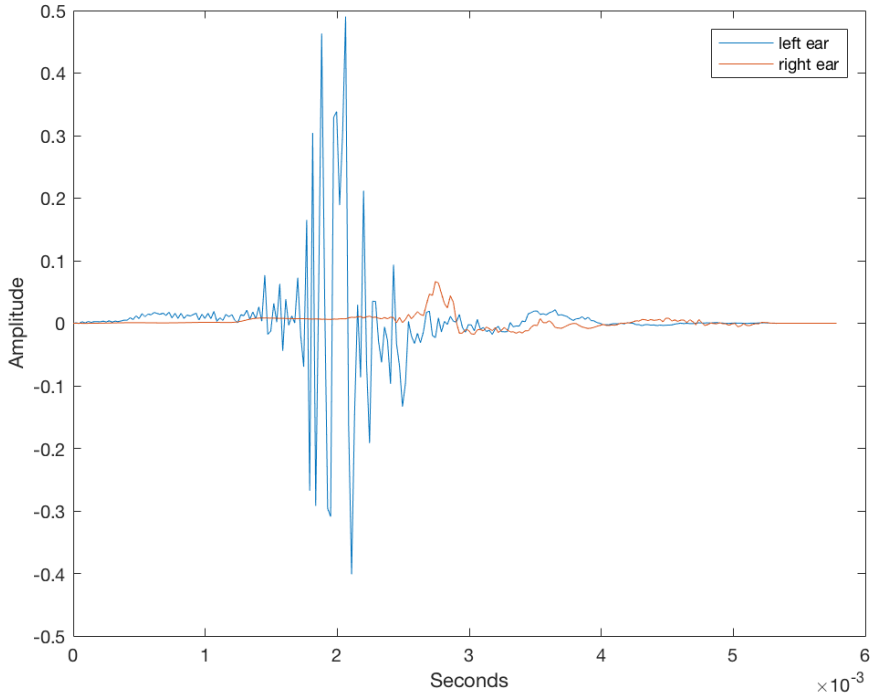


FIGURE 2.13: HRIR showing the difference in timing and amplitude between a left (blue) and right (orange) ear for a sound source positioned directly to the left of the listener. From SADIE database [58].

Binaural synthesis techniques are based on the concept of accurately reproducing the interaural and spectral cues present in a listener's measured HRTF. It is the belief that if these cues can be accurately replicated and subsequently played back to the listener in a manner that closely follows the original binaural listening experience, then they will hear spatialised sounds equivalent to a real world listening environment. A common method is to generate a set of filters based on pairs of HRTF measurements, that can then be applied via convolution to unaltered, monaural, sounds to impose the same spectral changes characteristic of a specific sound source position, relative to the listener [56]. The filtered signal is usually played back over a pair of headphones to ensure that the signal intended for a specific each ear is only heard by that ear [61]. If standard stereo loudspeakers were used for playback, then the signal intended only for the left ear would interfere with that for the right, and vice-versa, providing non-representative spatial cues. A technique called cross-talk cancellation exists to nullify this interference in stereo loudspeaker renders of binaural audio by using a set of filters in an attempt to

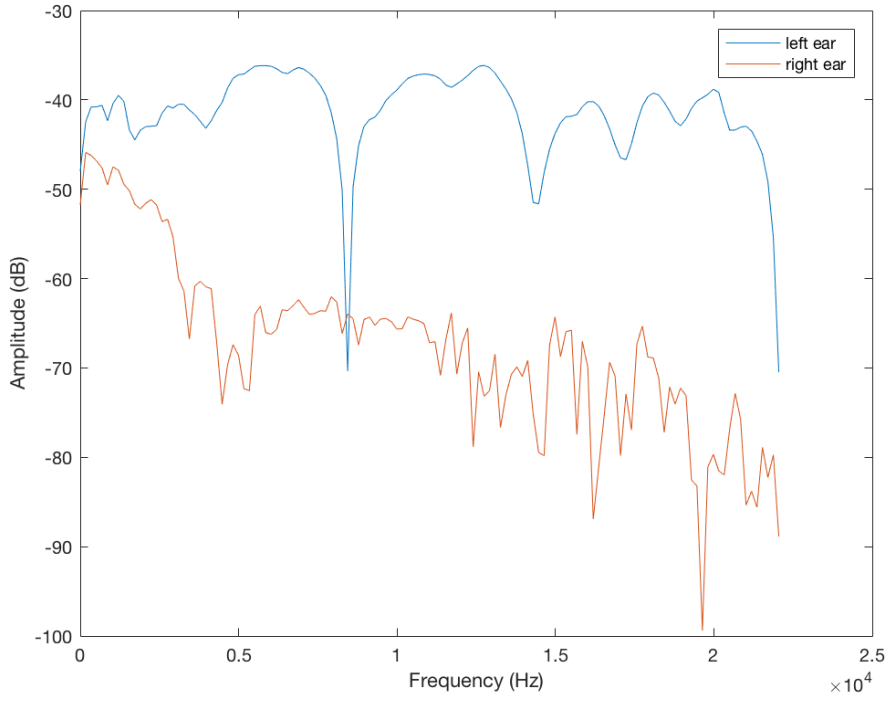


FIGURE 2.14: A HRTF plot for a left and right ear generated by performing a Fourier transform on the HRIR presented in Figure 2.13. The notches in the left ear plot (blue) are characteristic of the spectral cues caused by a sound interacting with the pinna. The altogether lower amplitude of the right ear plot (orange) illustrates attenuation of frequencies due to the head-shadowing effect.

separate the two audio channels as they are transmitted through the air to the listeners ears [62]. However, this can also introduce undesirable colouration to the timbre of the sound [63].

As stated previously, in order to gather a robust HRTF dataset, HRIRs should be measured in controlled conditions and specialised equipment, such as a set of in-ear microphones, is required. It is therefore difficult to obtain unique HRTF sets from individual listeners due to these practical limitations [64]. There are two main methods designed to overcome these issues, the first is to use an anatomically average dummy head, torso and ears in place of a real listener [65, 66], such as the KEMAR mannequin developed by G.R.A.S (see Figure 2.15). Such measurements are freely available in on-line HRTF databases such as CIPIC [67], SADIE II [58] and Listen [68]. The second method involves averaging the spectral content of different HRTF datasets taken from a number of subjects [56]. However, due to the uniqueness in the dimensions of the pinnae between individuals, there can be large differences in HRTF measurements obtained for the same sound source position. This includes notches appearing at different points in the frequency spectrum between listeners, as well as amplitude mismatches [64]. There is therefore a trade-off in the practicality of gathering

FIGURE 2.15: A KEMAR mannequin developed by G.R.A.S commonly used to generate generic HRTF measurements [69].

measurements and the consistency in simulated interaural and spectral cues across a wide range of listeners.

The use of HRTFs taken in anechoic conditions for binaural synthesis can result in the sound played back via headphones to be perceived to emanate from inside the head [56]. This is because, as stated in Section 2.5, the hearing system uses cues such as the precedence effect and D/R ratio to infer the direction and distance of sound sources. Anechoic environments are, by definition, non-reflective, and as such sound is not able to interact with the room in any way and thus these cues are lost. One way to overcome this is to gather a set of binaural room impulse response (BRIR) measurements for the desired positions around a listener. A BRIR captures the acoustic behaviour of the room, as well as the amplitude, timing and spectral cues used by the hearing system [70]. The measurement is taken using the same techniques outlined previously for obtaining HRIRs, the main difference being that it is not done in an anechoic environment. Utilising BRIR processing has been shown to improve externalisation of sound sources for headphone based playback [70, 71]

2.7 Summary

This chapter presented some of the fundamental theories needed to understand spatial hearing. This is important to consider since the way in which travelling sound waves interact with the surrounding environment, and the anatomy of the human body, play a major role in how sound is perceived spatially. The timing, amplitude and spectral cues used by individuals for directional hearing have been introduced, as well as the way in which these cues can be replicated over loudspeaker and headphone based systems to produce spatialised listening material of relevance in interactive gaming scenarios. For headphone playback, it is suggested that listening material should be processed using HRTF measurements to generate a spatialiased effect, or BRIR measurements if it is desirable to externalise sound sources.

Chapter 3

Multichannel Audio Playback

The term *multichannel audio* is used in reference to a collection of rendering techniques designed to present sound to a listener from multiple directions. In general, the aim of such techniques is to enrich a listener's experience of media content by promoting a sense of involvement. The involvement could come through something like feeling a sense of presence in a virtual setting, or by allowing a listener to become informed of narrative details through the use of specifically placed audio cues. Most video games are created such that they require an element of interaction from the player, meaning that the player has an active role in determining how and when in-game events, like sound effects, occur. The interaction from the player makes video games well suited to the benefits of multichannel audio. Spatialised sound cues can be used to fully envelop the player in audio, creating immersive virtual sound environments that dynamically react to player input. A large portion of the action can take place off-screen, either behind or to the sides of the player's viewpoint [72]. Audio cues can therefore be used to influence the player's actions by guiding them towards the next narrative event/objective or warning them of impending threats, potentially reducing the amount of visual information needed on-screen. From this, it is not unreasonable to think that video games enhanced with multichannel audio could make them altogether more engaging and offer clear tactical advantages for the player.

This chapter will introduce some of the main multichannel rendering systems available to general consumers, including both stereo and surround-sound. These systems are commonly used to play back multichannel audio for various types of multimedia content in both the home and larger viewing environments, such as cinemas. In general,

video game audio systems are those that are experienced in a home environment. Also, common terms used to describe the sensations a listener might experience whilst using a multichannel listening system are identified as a list of spatial attributes. These attributes form the foundation for the subjective questionnaire used to rate multichannel audio quality in some of the experimental work presented towards the end of this thesis.

3.1 Mono

Mono is the use of one audio channel to transmit sound to a listener, and as such is not considered to be a multichannel listening format. The format is not well suited for spatial playback because the perception of the majority of directional cues is reliant on the timing, amplitude and phase differences between the listener's two ears. For recorded audio this can only be replicated over at least two audio channels. It is possible to convey some spatial characteristics, such as distance, by manipulating the volume of sound sources and through applying DSP effects such as filtering, although the fact that phantom imaging and amplitude panning (see Section 3.2) between multiple channels cannot be achieved is a somewhat limiting factor [73]. Although there is scope to mix audio for a higher number of channels (namely stereo or surround-sound which are introduced later) some systems still exist for mono transmission such as those used by a number of FM and AM radio broadcasters [74]. It is therefore common for stereo or surround-sound audio mixes to be tested for mono compatibility, such that it can still be played back over the majority of audio playback systems.

3.2 Stereo

As a listening format, stereo reproduces audio over two channels to give a listener the impression that sound is emanating from multiple directions across a horizontal plane, known as the stereo panorama [75], which lies between the two loudspeakers (or headphones) used for playback. The two channels are divided into a Left (L) and Right (R), and as such are intended to be output (either over loudspeakers of headphones) from those respective positions relative to the listener. Stereo is a popular format and is used in the majority of music/audiovisual content, and has the highest user base amongst video game players [17]. It is important to note that in some cases stereo, as an

abbreviation of stereophonic, can refer to the use of more than two audio channels for extended multichannel playback [46]. For this thesis, stereo will be used in reference to the listening format consisting of a left and right channel only. Surround-sound will be used as a blanket term for any rendering format with more than two channels.

Using a separate left and right channel arises from the fact that humans perceive sound binaurally, i.e. with two ears. This has provided the basis for stereo recording and reproduction since the early 20th century with experiments conducted at the Bell Telephone Laboratories and by Alan Blumlein [76]. A microphone technique established by Blumlein in the 1930s [77] is an early example of stereo recording, in which a coincident pair of figure 8 polar pattern microphones are placed 90° apart to imitate the directionality of the left and right ears [78]. Ideally the microphone capsules would occupy the exact same point in space, but in reality they are placed as close together as possible due to the microphone housing, hence the term 'coincident'. The Blumlein stereo technique is illustrated in Figure 3.1. When recording, for example, a musical performance with multiple musicians, those positioned closest to the left microphone will be recorded at a higher level in that channel than those closer to the right, and vice-versa. This mimics the iteraural differences (introduced in Section 2.6) individuals use to infer the direction of a sound source in an environment [40]. The two recorded channels are then played back from two loudspeakers positioned to the left and right of

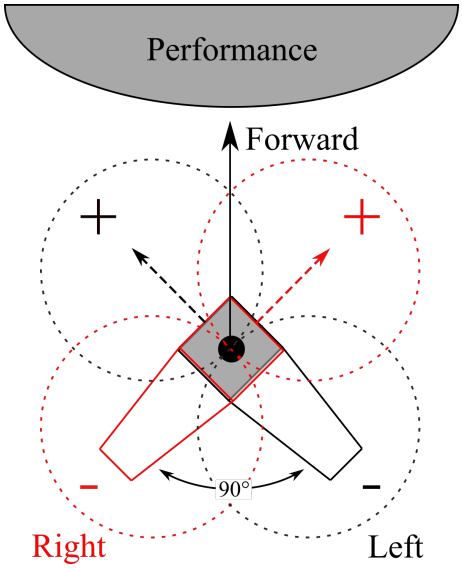


FIGURE 3.1: Illustration of a coincident pair of figure 8 microphones spaced 90° apart used in the Blumlein stereo recording technique. The red microphone represents the right ear of a listener and the black represents the left.

the listening position, giving the impression of a spatial audio experience.

Sounds that have not been recorded using a stereo technique (i.e. mono sounds) can also be positioned within the stereo field through a process of amplitude panning. The technique involves manipulating the relative amplitude of the signal across two audio channels to give the illusion of a phantom (or virtual) sound source emanating at some point between the two loudspeakers used for playback [79]. This can be used create the impression of movement between the two audio channels, whilst also separating sound effects within the stereo image. The ratio of gain values (g_1 and g_2) between the two loudspeakers shown in Figure 3.2 is calculated according to one of the possible sine panning laws given by Bauer [80]:

$$\frac{sin\theta}{sin\theta_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{3.1}$$

where θ is the perceived angle of the virtual source and θ_0 is the angle of the loudspeakers relative to a listener facing forward at 0°. If the listener's head is to be more formally considered then it is suggested that replacing the *sin* term with *tan* in (3.1) will provide more consistent imaging [81]. The actual gain values with a constant loudness can then be derived using:

$$\sqrt{\sum_{n=1}^{N} g_n^2} = 1 \tag{3.2}$$

where N is the number of audio channels, or loudspeakers. When using two loudspeakers, it is widely accepted that the angle between two loudspeakers should not exceed 60° if stable imaging is to be preserved, where $\pm 30^\circ$ relative to a forward facing listener is recommended [46], as illustrated in Figure 3.2. The wider the angle between the two loudspeakers, the more unstable the phantom imaging becomes, in that sounds are perceived to 'pull' towards the closest speaker relative to the intended position of the sound source, creating perceptual 'holes' in the stereo image. In the context of screen media content, this can cause disparity between audio and visual feedback, having potential implications on the user experience. Imaging can be further improved by

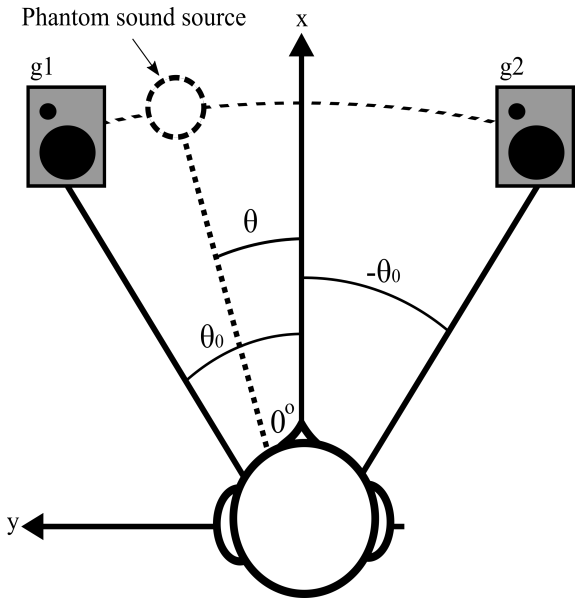


FIGURE 3.2: The relationship between the desired angle of a phantom sound source (θ) and two loudspeaker angles ($\pm\theta_0$) [81].

placing a centre loudspeaker, in-between the pair, at 0°, due to a narrower angle between adjacent loudspeakers. The spatial effect is also most stable when the listener is positioned within the sweet-spot, which is the central position between the two loudspeakers. The sweet-spot is illustrated by the position of the head in Figure 3.2. If the listener is not within the sweet-spot then the phantom imaging will be compromised in that it will be perceived to be closer to one loudspeaker than the other [82].

Although it has been noted that stereo has the highest user base amongst gamers, it is also important to consider that people may not be able to configure their stereo setup based on the recommendations outlined previously. Due to the differences between domestic living spaces, it may not always be possible to place loudspeakers at $\pm 30°$, or for a listener to sit in the exact sweet-spot. This therefore presents a challenge for multichannel sound content creators, as in a potentially high number of cases the stereo image may be lost.

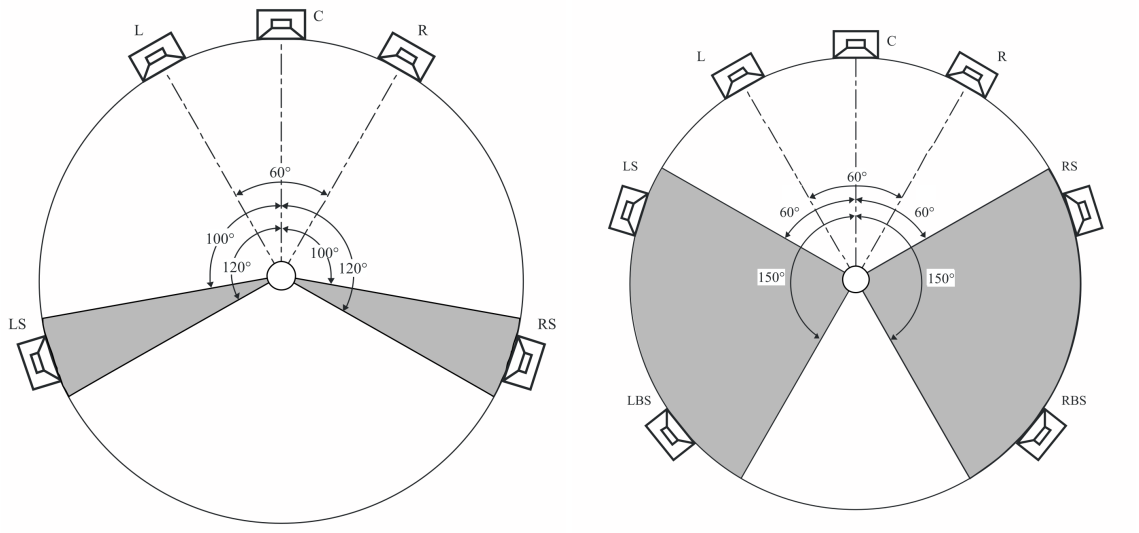
3.3 Surround-sound

One of the shortcomings of loudspeaker stereo is that imaging can only really be done successfully in the frontal quadrant, i.e between the left and right loudspeakers positioned to the front of the listening space. This makes it difficult to place any sound sources behind or to the sides of the listener. Surround-sound formats expand on stereo

by adding additional loudspeakers, usually to the rear and sides of the listening space, or in between the already established stereo pair. In most cases this retains the optimally placed stereo pair of $\pm 30^{\circ}$. This was the case with the original Dolby Stereo [83] format developed for large scale cinema sound in the 1970s which was later adapted for home cinema environments as Dolby Surround in the 1980's [84]. The format implemented an additional centre channel, reserved for dialogue, and one rear channel usually used to emphasise ambient film effects, such as rain and wind, and occasionally for discrete sound source placement [85].

By far the most popular commercial surround-sound formats for multimedia (at the time of writing) are 5.1 and 7.1 surround sound. The naming convention is derived from the notation $[x.y]$, where x represents the number of full-bandwidth channels used and y is the number of band-limited channels, usually reserved for low frequency effects (LFE). To that effect 5.1 surround-sound consists of five full range channels and one LFE, with 7.1 comprising of an additional two full range channels. Figure 3.3 shows the respective positioning of loudspeakers that are recommended for surround-sound listening, as suggested in ITU-Recommendation BS: 775 [86]. As with stereo, the best auditory imaging is perceived when the listener is sitting within the sweet-spot, i.e. the central point where all the loudspeakers intersect. Both formats comprise of a centrally placed loudspeaker (C in Figure 3.3) that is generally used in film to separate dialogue from the main soundtrack. This more closely matches the on-screen position of performers whilst also adding clarity. The centre channel can also be used to improve the resolution of frontal phantom imaging between the left/right stereo pair [87], by providing a physical source for sounds intended to emanate from directly in front of the listener. The use of the centre channel can vary widely between different video games and is discussed further in Chapter 4.

The surround and rear channels (RS, LS, RBS and LBS in Figure 3.3) are well suited for ambient sound effects, especially in cinema, where the importance of accurate sound source placement is often compromised in favour of maximum audience coverage. Sound sources can be placed around the listening space using the same amplitude panning principles discussed in Section 3.2, by manipulating sound source amplitude between pairs of adjacent loudspeakers located around the listening space. However, there has been dispute concerning the consistency of stable sound imaging when using



(a) 5.1 Surround-Sound (b) 7.1 Surround-Sound

FIGURE 3.3: Standard loudspeaker placement for both 5.1 and 7.1 surround-sound listening as suggested in ITU-Recommendation BS: 775 [86].

the loudspeaker placements for the rear and surround channels. The angles between the non-stereo and centre channels in both 5.1 and 7.1 surround-sound exceed the recommended angle of 60° and hence imaging becomes unstable and inconsistent, especially when considering lateral sound sources. Cabot [88] assessed the localisation of both rectangular and diamond quadrophonic (4 loudspeaker) systems, finding phantom imaging to be most stable in the front quadrants but very unstable to the sides and rear of the listener. This result is further emphasised by Martin et. al [89] who considered image stability in a 5.1 surround sound system. Their results suggest phantom imaging is both reliable and predictable using the front three loudspeakers but is highly unstable when a sound at a position greater than 90° relative to a front facing listener is desired. Theile and Plenge [90] suggest that for more stable lateral imaging, sound sources intended to be perceived at $\pm 90°$ relative to the listening position should be represented by a real sound source i.e. a loudspeaker. They propose an equally spaced arrangement of six loudspeakers to get a suitable 'all-around' effect. This configuration was extended by Martin et. al [91] to an equally spaced octagonal array with a front centre speaker placed at 0° relative to the listener. The array was found to give relatively stable imaging around the listening space for amplitude-based panning algorithms.

It is however important to consider that consistent sound imaging at every point around the listening space might not be important for many consumers. The side and rear channels used in modern surround-sound systems will give the impression that sound is emanating from around the listening space, even if this information isn't entirely accurate to what is happening on-screen. This in itself might provide enough spatial information for most casual viewers. In addition to this, there is an existing infrastructure for both home and large scale theatre sound systems based on 5.1 and 7.1 surround-sound. By replacing the channel/loudspeaker arrangement with, for example, an octagonal array the existing infrastructure would be disturbed, requiring the industry and consumers to adapt.

3.3.1 Surround-sound codecs

The multiple channels used for surround-sound playback are in most cases encoded into a single bit-stream before being transmitted to an end-user. This helps to reduce the amount of data needed for the audio when it is stored on either a physical format such as a Blu-Ray, or streamed over a network. The encoded information is then decoded back to the original surround-sound channels as and when it is necessary, usually by the device used to play the content or by some external system. Various codecs exist for this purpose and are most notably developed by DTS and Dolby who specialise in audio systems intended for multimedia, such as video games and film. For modern systems, this process is done digitally using audio signals that are represented digitally using linear pulse code modulation (LPCM). In short, this involves converting an analogue signal into a bitstream (a continuous sequence of binary code) consisting of only 1s and 0s, i.e a pulse. The conversion is done over three steps comprising of sampling, quantisation and encoding, as illustrated in Figure 3.4.

The amplitude of a signal is first sampled at equal intervals over time. The rate at which these samples are taken is defined by the sampling frequency (or rate), usually in Hz or kHz, where a higher frequency will result in more samples being taken over one second, thus retaining more of the original signal. The example given in Figure 3.4a shows a sampler working at 8kHz, i.e 8000 samples are taken every second. The sampling (or Nyquist-Shannon) theorem establishes that the sampling rate should be at-least twice that of the highest frequency captured to ensure that as close to the original signal is

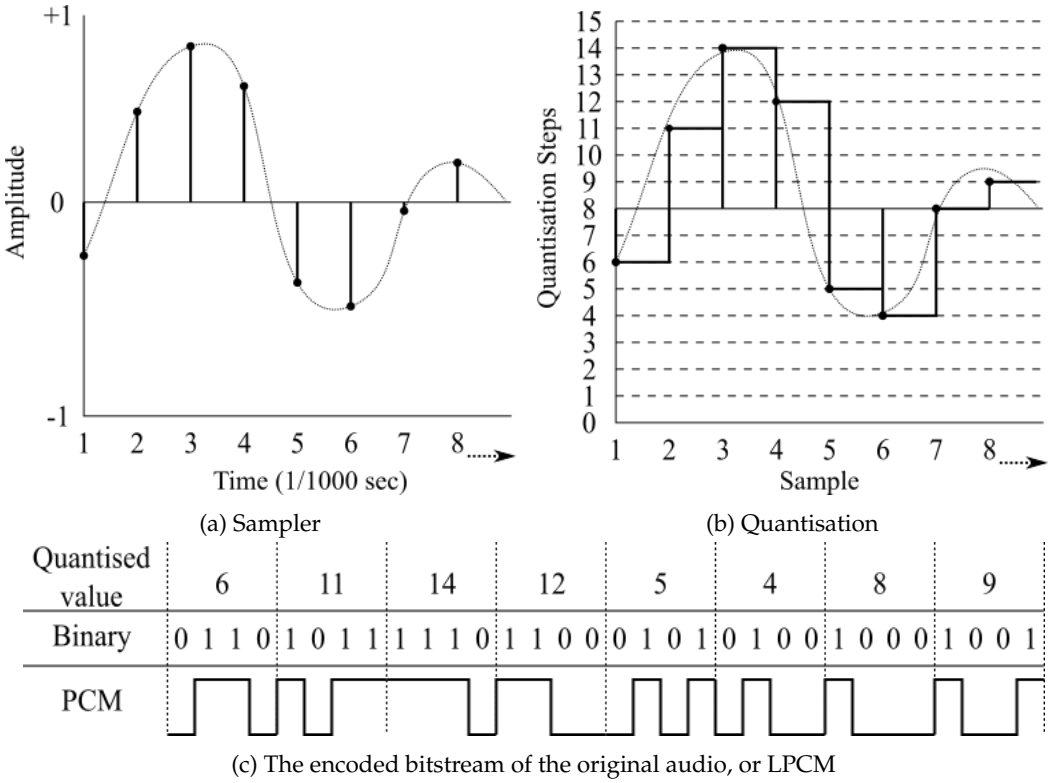


FIGURE 3.4: An illustration of an analogue audio signal being sampled (A), quantised (B) and encoded (C) to a digital bit-stream (LPCM).

restored through reconstruction of the samples [92]. If the sampling frequency is too low then the signal may become aliased [32]. Signal aliasing occurs when a different waveform to the original is constructed from the samples obtained in the sampling step [93]. Ensuring that the sampling frequency adheres to the sampling theorem will mean that the the original signal can be reconstructed from the recorded samples, and therefore no aliasing will happen.

The sampled amplitudes of the signal are quantised to discrete, linearly spaced values so that they can be represented as binary code. The bit depth determines how many steps the original signal will be quantised to, and how many bits of binary data (i.e how many 1s and 0s) each quantised value will be represented by [32]. The number of steps is equal to 2^n, where n is the bit depth. For example, in Figure 3.4b the chosen bit depth is 4, therefore there are 16 (2^4) steps of quantisation, each of which is converted into 4 bit binary at the encoding stage (see Fig 3.4c). As the sample rate and bit depth increase, so does the amount of storage space needed for each audio channel, hence the need to encode surround-sound audio into a more manageable bit-stream. Codecs provided by

DTS and Dolby perform a further encoding step that compresses the multiple audio channels before being stored on the chosen data format. The exact methods for doing this are commercially sensitive, therefore there is little documentation on the actual procrss. Modern codecs for up to 7.1 surround-sound include DTS-HD Master Audio [94] and Dolby True HD [95], both of which are lossless. This means when the encoded bit-stream is decoded, it is the same as the original LPCM signal for each channel. For gaming, DTS Digital Surround and Dolby Digital are examples of codecs that are capable of decoding game audio in real-time, meaning that sound sources can be panned around the surround system relative to user input.

3.3.2 Stereo Mix-down

Surround-sound material can be presented over any regular stereo system, at the expense of fuller spatialisation, by down-mixing non-stereo channels. The surround and center channels are attenuated and then combined with the front left and right to ensure no sound effects/cues are lost in the process [86]. For example with 7.1 content, the same channels used when listening over a loudspeaker array would be the ones down-mixed to two channel stereo. In reference to Figure 3.3, the left/right surround (LS/RS) and left/right back surround (LBS/RBS) are attenuated by 3dB and sent to the respective front left and right channels. For Dolby systems, a 90° phase shift is also applied to the surround channels in order to more easily generate the down-mix [96]. The centre channel is also attenuated by 3dB and then transmitted equally to the front left and right. For 7.1 surround-sound this is expressed in ITU-R BS: 775 [86] as:

$$\begin{aligned} L_D &= L + 0.707C + 0.707Ls + 0.707Lbs \\ R_D &= R + 0.707C + 0.707Rs + 0.707Rbs \end{aligned} \tag{3.3}$$

where L_D and R_D are the down-mixed left and right channels, respectively. The method is beneficial to content creators, as it negates the need to generate separate mixes of what is essentially the same audio material. It also provides an easily implemented solution for headphone presentation. The down-mix is usually performed on whatever system is being used for playback. As long as a compatible codec is provided, these devices

include television sets, DVD/Blu-Ray players, games consoles and external amplifiers. The codec used to compress the surround channels will also encode some metadata into the bit stream for down-mixing purposes. In the digital to analogue conversion (decoding) process at the playback stage, this metadata is used by the system to determine whether the user is listening over a surround-sound or stereo set-up [97]. If a stereo system is detected, like the default loudspeakers on a TV, then the down-mix of surround channels will occur automatically. Some devices also allow users to manually switch between a stereo down-mix and the original surround channels.

3.4 Virtual Home Theatre (VHT)

Virtual Home Theatre (VHT) systems offer another headphone based approach for surround-sound listening. Loudspeaker channels are virtualised by processing the individual audio channels with HRTFs [46], like the binaural synthesis systems introduced in Chapter 2. For each surround channel, HRTF pairs are gathered by taking impulse responses at the ears of a real listener, or a dummy head, relative to loudspeaker positions conforming to either 5.1 or 7.1 surround-sound standards [59, 98]. The HRTF measurements for each loudspeaker position are then convolved with the corresponding surround-sound channel. In theory, this allows listeners to experience all the individual channels of a surround-sound system, over a pair or regular stereo headphones. It also means that the panning effects between different channels are not compromised in the same way as a stereo down-mix. This is an altogether more convenient solution for surround-sound listening as it negates the need for specialist equipment, like amplifiers and loudspeakers, and a dedicated listening space. However, in reality a VHT system is rarely equivalent to the physical loudspeaker counterpart. Various subjective sound quality studies have shown that if possible, it is preferable to use a physical surround-sound system comprising of suitably positioned loudspeakers rather than a VHT system [99–101]. Examples of consumer VHT systems include the Turtle Beach i60 headset [102] and Razer Surround software for PC [103].

3.5 Ambisonics

The term Ambisonics refers to a multichannel encoding technique, developed by Gerzon [104], based on the decomposition of a sound scene using spherical harmonics. Spherical harmonics are a set of functions that represent areas on the surface of a sphere [105]. The format offers an interesting alternative to surround-sound codecs, since Ambisonically encoded audio can, in theory, be decoded to any number of loudspeakers in any configuration. B-format is the name given to the collection of channels used to transmit the encoded soundscape before being decoded to some pre-defined loudspeaker arrangement [46]. Individual sound sources, like recorded samples, can be encoded directly into B-format using an Ambisonic panner. The encoding process effectively places sound sources at different points on a unit sphere surrounding the listener [106]. This can also be done in real-time, meaning that the encoding can react dynamically to user interaction. Examples of such panners are the *ambix* VST plugin for use in the digital audio workstation (DAW) Reaper [107], or in the *Spatialisateur* (Spat∼) object library for Max/MSP [108]. Using first order B-format as an example, the process can be expressed mathematically as follows [109]:

$$\begin{aligned} W &= input \\ X &= input \times \cos(\theta) \times \cos(\phi) \\ Y &= input \times \sin(\theta) \times \cos(\phi) \\ Z &= input \times \sin(\phi) \end{aligned} \tag{3.4}$$

where *input* is the mono sound source to be encoded, θ is the azimuthal (horizontal) angle of that sound source and ϕ is the elevation (vertical) angle, relative to a fixed, front-facing position [110]. Therefore θ and ϕ correspond to the angles at which the sound source is intended to be percieved from, after the B-format has been decoded and rendered over loudspeakers. As the Ambisonic order increases, so do the number of channels used to encode the audio material. When this is the case a weighting factor is applied to the individual signals by the SN3D normalisation scheme. The normalisation ensures that the peak amplitude of each encoded source will not exceed that of the zeroth order signal (i.e. the W channel) [111]. When the Ambisonic order is higher, the

soundfield is decomposed using a greater number of spherical harmonics. This results in a perceptual increase in the spatial resolution of the recorded soundfield, meaning that accuracy in localisation between different sound sources is improved [112–114].

One benefit of B-format is that transformations, such as rotations, can be applied to the encoded soundscape with relative ease using rotation matrices. This is beneficial when considering the VHT systems discussed previously. When played back over headphones, the virtual loudspeakers will follow the listener's head movements which is not representative of a real listening environment where loudspeakers should be stationary. By tracking the listener's head movements, compensatory rotations can be applied to the virtual soundstage in order to stabilise the positions of the virtual loudspeakers [56]. A soundscape encoded into first order B-format can be rotated horizontally using the rotation matrix presented in (3.5) from [106]:

$$\begin{bmatrix} W' \\ X' \\ Y' \\ Z' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos\theta & -sin\theta & 0 \\ 0 & sin\theta & cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} W \\ X \\ Y \\ Z \end{bmatrix} \tag{3.5}$$

where W',X',Y' and Z' are the rotated versions of the B-format channels W,X,Y and Z by angle θ (the horizontal angle of the listeners head in degrees). In the case of a VHT system, dynamically performing this transformation will help to give the impression that virtual loudspeakers are located in fixed positions, regardless of headphone orientation and, hence, resulting in a stable, 'fixed' soundscape.

3.6 Spatial Attributes

According to Berg and Rumsey, spatial attributes are terms used to describe the 'three-dimensional nature of sound sources and their environments' [115]. These attributes often form the foundation for subjective listening tests with the purpose of determining how well a multichannel listening system, such as surround-sound, can convey the three-dimensional (or spatial) characteristics of audio to a listener. Comprehensive lists of spatial attributes, with descriptors, are given in a number of publications concerning the assessment of listening systems. Most notably this includes the *Spatial Audio Quality Inventory* [116], as well as work by Le Bagousse et al. [117, 118],

Rumsey [73] and, Bech and Zacharov [119]. There are similarities in the attributes identified between these works, therefore the most frequently mentioned are discussed in this section with a focus on how they might be utilised in an interactive game environment. These attributes are also used in the experimental work presented in Chapters 6 and 7, to determine how multichannel game audio is perceived by a player.

3.6.1 Distance

The distance of a sound source refers to how far away (or close) from the listener it is perceived to be. As introduced in Chapter 2, an individual can infer the distance of a sound source based on the relative loudness, as determined by the inverse square-law, and the ratio between the direct and reverberant sound. In addition to this, high frequencies are also attenuated as a result of air absorption [120]. Over a playback system, distance can therefore be simulated by manipulating a signal's amplitude, spectral content and reverberant energy. Effective simulation of sound source distance is necessary for game audio, as it can provide the player with a sense of scale in regard to a virtual environment. It can also be used as a tool to inform players of how far away certain in-game objects and objectives might be, thus potentially informing player decisions.

3.6.2 Localisation

Localisation is an individual's ability to infer the direction from which a sound source is emanating in terms of the relative azimuth (lateral position) and elevation (vertical position). As defined in Chapter 2, the interaural differences of a sound pressure wave between the two ears, and the spectral changes caused by the size and shape of the pinna are important cues in localisation. Effective simulation of source localisation is perhaps one of the most useful audio cues in games since in-game sounds can often occur to the sides or behind the player viewpoint [72]. Using a surround-sound system, sound cues can be output from multiple positions around the listening space, meaning sound sources that are not represented on-screen can still be heard. This has potential benefits for enhancing player immersion and providing tactical advantages, since a player can be made more aware of the surrounding events in the virtual environment [2, 3]. The game

designer should ensure that each sound source is placed accurately and appropriately in the virtual space as to make the direction obvious to the listener [116].

3.6.3 Depth

In much the same way that an artist would utilise perspective in a painting to provide dimensionality, sounds with depth will be perceived to recede from the listener providing as sense of scale within a virtual environment [116]. As illustrated in Figure 3.5 depth can be thought of as the front/back definition of a sound source (the source depth), or a group of sound sources (the ensemble depth) [73]. When simulating depth, it is the perceived distance between the front and rear most points of the presented sound image that cause the sensation [121]. Depth should not be confused with distance perception, as it relates to how the relative distance between multiple sound sources, heard at the same time, can help to create a collection of sounds that are perceived to be both close to and far away from the listener. An example of this in a gaming environment would be for a player to be placed in the midst of a large crowd where the voice of a closely situated bystander can be clearly heard but the player is still able to hear the general bustling of the rest of the crowd fading in the distance. The effect can be simulated by manipulating the amplitudes of multiple sound sources relative to each other.

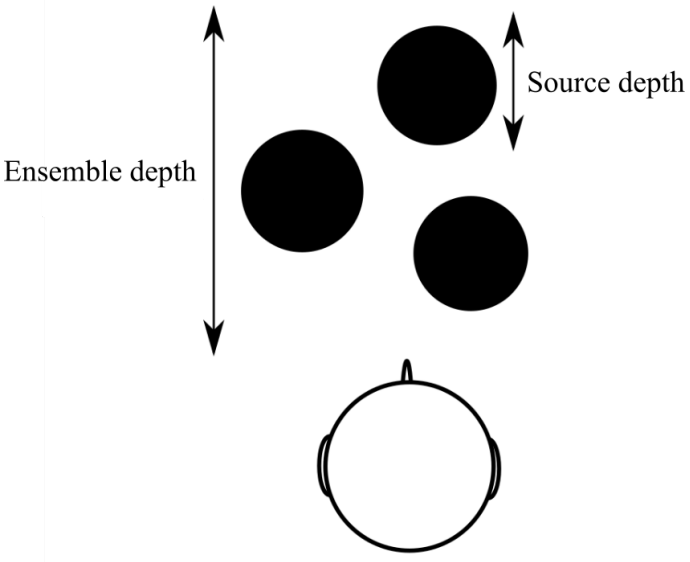


FIGURE 3.5: An illustration of the relationship between individual sound sources depth and the overall (ensemble) depth of the sounds presented in the environment, as adapted from [73].

3.6.4 Width

The width of a sound source is defined by how large a space it is perceived to occupy in the horizontal direction [122], as determined by early lateral reflections arriving at the two ears approximately 80-100ms after the direct sound [123]. It is important to note that in highly reverberant environments, a sound may appear to occupy a larger space because an increase in early lateral reflections will result in a perceptual widening of the source [43, 124]. In a similar manner to depth, source width relates to groups of sound sources as well as individual ones. According to Rumsey [73] there are three main types of source width: *individual*, *ensemble* and *environment*. Although these descriptors are derived from musical terms, their applicability to virtual soundscapes is clear by replacing groups of instruments with groups of in-game sound effects. *Individual source width* refers to the separate auditory components of a soundscape and their perceived left-right extent. A group of sounds that are considered to make up one single, often larger, entity (for example the engine hum, wheels and brakes of a moving car) have an *ensemble width*. *Environment width* refers to the presented space and how narrow or wide it is perceived to be based on its reverberant energy. These three types of source width are illustrated in Figure 3.6.

Aesthetically, the effective and appropriate use of source width in video game audio provides players with a better sense of scale regarding the presented sound-scenes and their contents. Width can be simulated for video game players over loudspeakers by the appropriate mapping of a single sound source to multiple speakers at the same time. For example a large sound source, with individual source width, might be heard from two or three adjacent speakers simultaneously, giving a sense of its size.

3.6.5 Envelopment

In the context of audio, envelopment is the extent to which a listener feels surrounded by the combined sounds in a physical, or virtual, space [116]. This attribute not only relates to the acoustic properties of the space, i.e. the early and late reflections, but also the way in which the directionality of these reflections are presented to the listener over loudspeakers or headphones. If possible, each sound source in the sound scene should have its own unique early reflection properties to give more natural and believable room simulations [125]. This degree of accuracy may not always be possible due to the

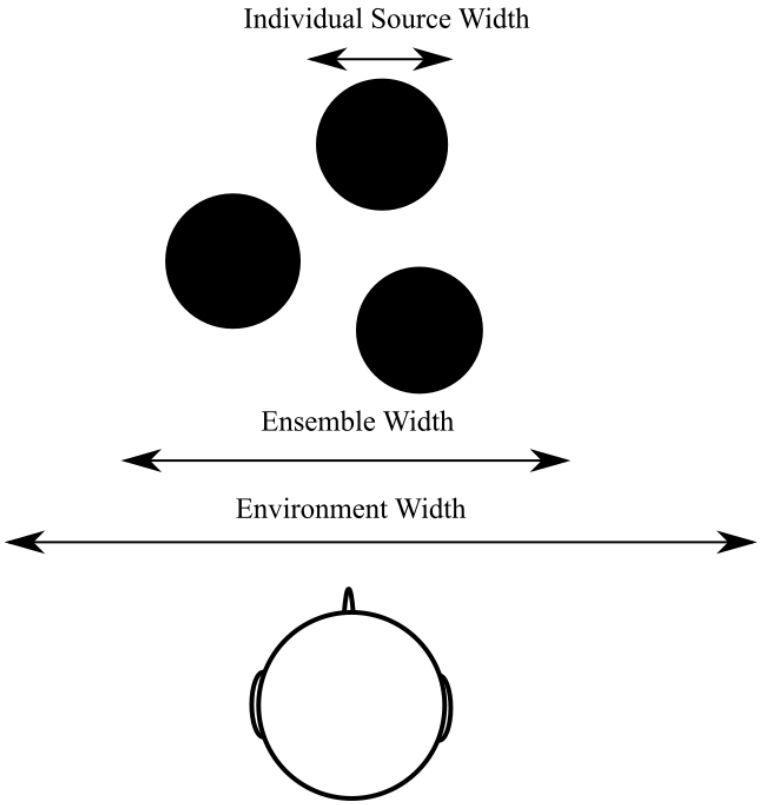


FIGURE 3.6: The relationship between individual, ensemble and environment sound source width, as adapted from [73].

potential resource load such a system may require, however the addition of loudspeakers to the rear and side of the listener helps to further emphasise a sense of envelopment. In creating a more enveloping listening experience, through dynamic reverberation and surrounding sound effects, the listener will feel they are part of the virtual world, rather than outside of it [126].

3.7 Summary

This chapter has introduced some of the more common multichannel listening formats available to consumers, including stereo and variations of surround-sound. At the time of writing, 7.1 surround-sound is the system with the highest number of available channels used in the majority of video game content (as evidenced in the next chapter). Although 7.1 surround-sound cannot convey a fully three-dimensional audio rendering of a virtual environment, due to the lack of loudspeakers above and below the listening

space, the phantom imaging of sound sources that can be achieved between pairs of adjacent loudspeakers surpasses the ability of other commonly used systems, such as 5.1 surround-sound. The ways in which surround-sound content can be presented to individuals who do not have access to the necessary equipment have also been considered, most notably stereo down-mixes and virtual home-theatre (VHT) renderings for headphone listening. These options offer a more convenient approach to surround-sound listening, since only a pair of regular stereo headphones is required. However, in using such systems, the spatial elements of a surround-sound mix are often compromised due to the lack of physical sound sources (such as loudspeakers) surrounding the listener. For these reasons, the perceived quality of both physical and virtual 7.1 surround-sound will be considered in the experimental work in the latter parts of this thesis. It is of interest to investigate whether the shortcomings of stereo and headphone based systems are noticeable whilst a player is engaged in playing a video game, and if this will impact their experience in any way. The way in which the quality of multichannel audio might relate to a user's experience is discussed in Chapter 5. In order to gauge the perceptual quality of these multichannel playback systems, commonly used spatial audio attributes have also been identified and defined. These will form the foundation for the subjective listening tests presented throughout this thesis.

Chapter 4

Multichannel Video Game Audio

Thus far multichannel audio has been defined generally through examples of common listening systems for multimedia experiences. This chapter focuses more specifically on surround-sound systems, a subset of multichannel audio formats, that are commonly used in video game content. Firstly, a historic review of multichannel audio in gaming is given, based on a timeline of video games consoles and their audio capabilities. This is mostly made up of sources found on-line due to the lack of similar summaries in more formal literature. Observations of surround-sound implementation in modern video game content is then given in order to express some of the creative approaches to rendering and differences in comparison to the film industry. Specific game examples are given focusing on some of the main differences between games such as: the formats used, the use of the centre channel, music panning and camera perspectives. Considerations from here will also be used to inform game design decisions in the experimental work introduced in Chapter 8. The chapter ends with in-depth reviews of four games; *The Last of Us: Remastered*, *Alien Isolation*, *P.T.* and *Ratchet and Clank: Tools of Destruction*. These games are considered as potential stimuli for the experiments in Chapters 6 and 7, therefore the pros and cons of each are discussed in terms of a set criteria.

4.1 A History of Multichannel Audio in Gaming

Before delving into the potential impact multichannel audio may have on gameplay experiences, it is important to note the current state-of-the-art and reflect on some of the milestones in game audio up to this point. Some of the key developments are shown on

FIGURE 4.1: A timeline showing some of the key developments in multichannel video game audio for home gaming consoles.

the timeline in Figure 4.1. Fairly complex audio systems had been common in arcade and pinball machines for some time, contrasting to early home gaming hardware which often had very limited sound capabilities. An early example of interactive and dynamic audio in arcade machines was *Space Invaders* released in the late 1970s. As the player destroys more enemy spaceships, the speed at which the ships move towards the bottom of the screen increases and the tempo of the soundtrack also increases to reflect this. This is regarded as an early example of audio directly reacting to the actions of the player, hence creating a dynamic soundtrack, however it is interesting to note that the developer of *Space Invaders* admits that this feature was originally caused by an accident within the game's code [127].

The Magnavox Odyssey [128], which is often regarded as the first home gaming console released in 1972, didn't have any audio output at all and later machines from the mid 70's, such as PONG [129], were limited to monaural beeps that were only made possible by adapting the hardware that was already being used to run the gameplay [130] (i.e. there were no dedicated sound chips). The Commodore Amiga 1000 [131] is believed to be the first home gaming computer that could output two separate audio channels, offering the possibility for stereo playback. This was thanks to the Paula soundchip [132], which could also handle 8-bit digital audio, allowing games to move away from synthesised sounds, and instead towards employing recorded samples. Stereo still has the largest user base amongst players and is standard in almost all games [17].

As introduced in Chapter 3, imaging to the sides and rear becomes difficult due to there being only two, generally frontally positioned, channels of audio. Surround-sound was implemented in a handful of titles for the Super Nintendo Entertainment System [134] [135], making it the first example of a game console to utilise the Dolby Surround home theatre standard [84]. The technique extended the conventional stereo format through the addition of a surround channel used to drive loudspeakers to the rear of the listening space [136]. For gaming, this rear channel was often reserved for ambient sounds, such as weather effects, or music, much like in a cinema environment [137]. Notable games include *Jurassic Park* [138], *Vortex* [139], *Samurai Spirits/Shodown* [140], and *King Arthur's World* [141], a screen-shot of which can be seen in Figure 4.3. Up to this point, home theatre standards, like Dolby Surround, had only be used for film and potentially music. Therefore with an existing infrastructure already in place, it perhaps made sense to game

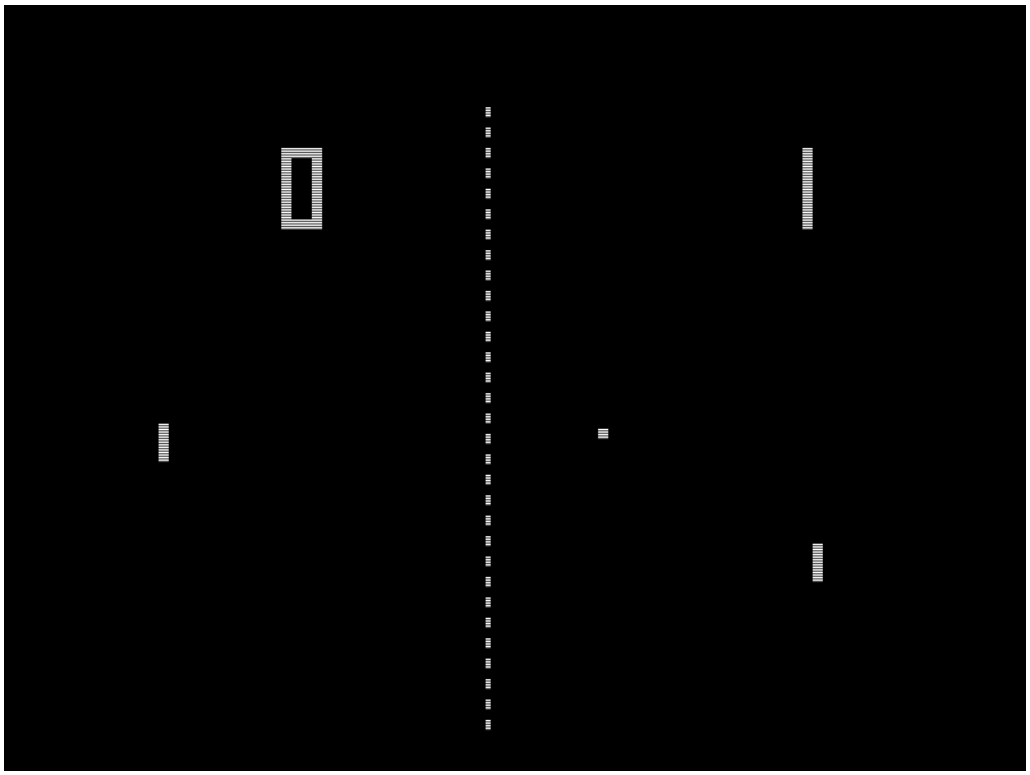


FIGURE 4.2: A screen-shot of Pong [133] developed by Atari. One of the first home video games/consoles to have audio.

content developers to make use of them to extend the game audio playback. This has been the trend with new surround formats and gaming since, where generally the latest generation of consoles will use the most popular home-cinema format of the time.

5.1 surround-sound was implemented for Sony's Playstaion 2 (PS2) in 2000, and was the first games console to do so [143]. The inclusion of a DVD drive meant that the console was able to make use of surround-sound codecs including Dolby Digital, Surround and Pro Logic II via a separate decoder (such as an A/V amplifier) connected to the console with an optical cable. However, discrete 5.1 surround-sound was rarely used during gameplay and was often reserved for pre-rendered cut-scenes and DVD-video playback due to the difficulties in encoding game audio to an appropriate surround-sound format using the PS2 hardware [144]. The DTS Interactive codec [145] was implemented in a handful of PS2 titles, such as *Grand Theft Audio: Vice City* [146] and *FIFA Soccer 2003* [147], allowing for real-time encoding of game audio, for 5.1 playback, during gameplay [148]. In-game 5.1 audio became standardised with the release, by Microsoft, of the original Xbox console in 2001. The Xbox hardware was developed such that game audio

FIGURE 4.3: A screen-shot from the video game *King Arthur's World* [142], one of the earliest games to make use of Dolby Surround. Certain instruments and sound effects in the game's soundtrack are panned to the rear channel, to give an all-around effect [137].

could be easily encoded or decoded for 5.1 playback in real-time using the then new Dolby Digital Live (DDL) codec [149]. Because of this, almost the entire back catalogue of games released for the Xbox are capable of full 5.1 surround-sound game audio playback, including titles such as *Halo: Combat Evolved* [150], *Fable* [151] and *The Elder Scrolls III: Morrowind* [152]. At the time of writing 7.1 surround-sound is standard in almost all games from core developers, with the Playstation 3 (PS3) [153] being the first console to offer this feature/functionality. The Playstation 4 (PS4) [154] and Xbox One [155] (part of the current generation of consoles at the time of writing (2018)) implement Dolby True HD [95] and DTS HD Master Audio [94], the current generation of 7.1 surround-sound codecs. Interestingly, the PS4 was the first home device to support the DTS HD Master Audio codec [156]. The way in which surround-sound is used in gaming, and how this might differ to conventions found in the film industry, is expanded upon in Section 4.2.

Although the focus in the current chapter is on console-based gaming systems (i.e. those systems defined by the current console generation) it is important to note that personal

FIGURE 4.4: A screenshot from *Halo: Combat Evolved* [157], originally released for the Xbox in 2001 and one of the first video games to allow for full 5.1 surround-sound game audio playback.

computers (PCs), running either Windows, Macintosh ot Linux based operating systems, are also a popular choice amongst video game players. PC systems are not limited in the same way as consoles in that internal hardware is interchangeable and therefore the system can be upgraded/improved at a much quicker rate. Dedicated PC sound cards compatible with surround-sound codecs have been available to PC gamers since before the introduction of the PS2 and Xbox in 2000 and 2001 respectively. It is difficult to say exactly when surround-sound was first used in PC gaming due to vast number of soundcards released by different manufacturers, however the *Soundblaster* series, made by Creative Technologies, have incorporated the option since the late 1990s [158]. Another difference with regards to audio for PC gaming is that headphone based virtual surround-sound methods (see Section 3.4) have been much more prevalent than with consoles. This is down to the Application Programming Interfaces (APIs[1]) often used in PC software and hardware development. Popular PC audio APIs include *DirectX* [160], *OpenAL* [161], *A3D* [162] and products by *Blueripple* [163], all of which have incorporated VHT for video game surround-sound based on HRTFs and binaural audio synthesis. Examples of PC games with the option include *Bioshock* [164] and *Doom 3* [165], however the use of such systems seems to have fallen out of fashion in recent

[1] An API contains pre-generated functions, command and code designed to allow interaction between different applications [159]. This allows developers to create different interactive systems without needing to write the code from scratch.

times due to lack of support from the developers and in modern operating systems. It is now more common for VHT rendering to be done separate to the actual game, by processing the surround-sound output using a 'black-box' hardware or software alternative. This includes the Turtle beach i60 headset [102] and Razer Surround software [103]. VHT approaches are beneficial because the user does not need access to a full loudspeaker array, and the 'black-box' approach means that the rendering does not need to be done on the game's output, but instead via an external system. The downside to this is that players will not have a VHT rendering by default.

As of 2017, Dolby Atmos [8] (originally a standard for cinemas) is the current state-of-the-art in multichannel game audio playback, being implemented in the PC versions of both *Star Wars Battlefront* [9] and *Overwatch* [10]. In these titles Dolby Atmos is most notable in the use of channels for height, to output sound from above the listener. Binaural processing for headphone listening is also becoming more popular with the prevalence of virtual reality (VR) systems like the HTC Vive, Oculus Rift and PlayStation VR. In a VR application the user perceives the virtual world through a head-mounted-display (HMD), rather than on a stationary television set/computer monitor. Headphone playback is therefore appropriate, because head-movements can be more easily tracked and applied to audio presented in this way.

It is interesting to observe that multichannel audio has been used in gaming for a relatively long time, first with stereo and then with more advanced surround-sound techniques. Clearly it is believed to be an important aspect of the overall game experience by developers, even though the majority of gamers don't actually have access to the systems/loudspeakers needed [17]. The next section will explore how the use of surround-sound in modern gaming content is used, and how this might impact the player experience.

4.2 Observations on Multichannel Game Audio

Multichannel audio in gaming offers an interesting opportunity for game audio mixing in that, for example, it can be used to envelope the player with audio, creating immersive virtual sound environments, or be used to alert the player to surrounding objectives. By playing through a number of different video games it became clear that multichannel

implementation can vary widely, suggesting there is no set framework for the way in which it can be used. Publications by Kerins [2, 3] provide some insight as to the practices of multichannel audio in gaming with a focus on 5.1 surround-sound. This section will expand on this work by providing a summary of surround-sound implementation in video games, with some modern (as of 2018) content examples. It is important to note that due to the vast pool of video game content available, it is not possible to provide details on all practices observed in gaming.

4.2.1 Formats

The majority of game content released for the current generation of home consoles, that being the PS4, Xbox One, Nintendo Switch and PC/Mac, will output some form of multichannel audio format, usually stereo or 5.1/7.1 surround-sound. As introduced in Section 4.1, surround-sound has been implemented in games since the 1990s, however, it is only with the more recent console generations that 5.1 or 7.1 surround-sound are common. It is rarely made clear on a game's retail box/artwork or on-line specifications as to what surround-sound formats are actually supported. Information will often be given regarding the implemented codec (either Dolby, DTS or both), but multiple loudspeaker systems are usually supported by the same codec. It is therefore often necessary to first test the audio output using an external A/V amplifier connected to the console to ascertain the supported loudspeaker layouts. This is important to keep in mind when short-listing appropriate content for user testing, since even if a game is released for the current generation and makes use of a modern codec, there is no guarantee it will actually output to all the channels in a surround-sound system.

The vast majority of end users will, however, listen to game audio using either television loudspeakers or headphones, rather than full loudspeaker surround-sound systems [17]. It is also rare for game audio to be rendered by default using a VHT system (see Section 3.4) for surround-sound listening over headphones [98]. Game audio content will often be down-mixed by the console to stereo so that it can be listened to over any standard stereo system, the internal loudspeakers of a television set, or headphones. Through observation, it seems that this is done automatically if the console is not connected to an external A/V amplifier. The PS4, for example, will default to stereo if the user is listening using the TV loudspeakers, but allows the user to change the audio rendering

in the system settings if they desire. In-game audio settings, usually accessed when the player pauses gameplay, are also common, with basic volume adjustment between audio groups such as dialogue, music and sound effects. Some audio settings, such as those found in *Grand Theft Auto V* [166] and *The Last Guardian* [167] will also allow the user to define the type of listening equipment they are using with examples being *Home Cinema, Headphones or TV*, meaning each has a specially tailored sound mix.

There are a few examples of VHT audio in gaming. Notably, this was done in the mobile game *Papa Sangre 2* [168]. The player receives no visual feedback and must navigate the game world using only binaural cues played back over headphones. This includes following a voice to objectives and avoiding enemy NPCs based on the sounds they emit. A more recent example is in *Hellblade: Senua's Sacrifice* [169], released by Ninja Theory in 2017 and winner of the 2018 BAFTA award for Audio Achievement [170]. The main protagonist suffers from psychosis, and as such whispering voices recorded using in-ear microphones are used to simulate the symptoms by creating the illusion of multiple disembodied voices surrounding the player [171]. It is interesting to note the different ways in which both games use binaural processing to influence the gameplay. *Papa Sangre 2* uses sound to guide the player and thus as a navigational tool, whereas *Hellblade: Senua's Sacrifice* uses the sound to further define the narrative, and elicit an emotional response from the player.

The Ambisonic format has also been used in a handful of games, most notably those developed by *Codemasters*, such as *DiRT* [172], although it has not been widely adopted as a gaming surround-sound format. Although there are not yet any standard sound design workflows for Ambisonic game audio, as of 2016, plug-ins to encode and decode Ambisonic game audio have been available in *Wwise* [173], a leading game audio middleware by Audiokinetic, therefore the format may become more common. Ambisonics is also used in virtual reality (VR) applications, an example being the *Resonance Audio* [174] plug-in developed by Google, designed to render B-format audio from a game engine (such as Unity) for binaural listening. In VR, it is necessary to constantly apply rotations to the surrounding soundscape such that audio reacts to the input of user head movements. As introduced in Section 3.5, this can be done fairly easily when audio is encoded in Ambisonic B-format, hence the wider use in VR over other games viewed using a stationary television set.

4.2.2 Perspective

In gaming, the term perspective refers to the point-of-view (POV) from which the player sees the virtual game world. The different perspectives considered in this thesis are first-person, third-person, top-down (or birds-eye) and side-views. The perspective used in-game usually defines the way in which sounds effects are panned around the loudspeaker array, and how these sounds relate to the virtual in-game objects seen by the player. This section will define some of the perspectives frequently used in video games, and how this might impact multichannel rendering.

Using first-person, the game is viewed as if through the eyes of the avatar controlled by the player. Many games from varying genres use a first-person perspective such as: *Call of Duty 4: Modern Warfare* [175], *Portal 2* [176], *Surgeon Simulator* [177] and *Minecraft* [178]. The first-person POV has become so common that first-person shooter (FPS) games are a definitive game genre. Multichannel audio rendering is fairly consistent between different first-person games (if a real-time codec is implemented), in that sounds tend to pan around the view-point, relative to the in-game object emitting the sound. Using recorded dialogue as an example, an NPC might communicate some important information to the player whilst first standing directly in-front of the viewpoint, and then whilst continuing to speak, walk around to the left-hand side of the POV such that they are no longer on screen, but can still be heard. If the game audio is being rendered to a 7.1 surround-sound system, the dialogue would first playback from one of the frontally positioned loudspeakers (i.e. left, right or centre) and then pan to the left surround channel, in-sync with the movement of the NPC. Panning like this will also occur if the player manually rotates or moves the viewpoint of their in-game avatar around sound emitting in-game objects. The use of surround-sound in first-person games enhances the embodiment of the character controlled by the player, giving the impression that sounds within the virtual game world react dynamically to external input.

The reboot of *DOOM* [180], released in 2016, uses the fact that the player is perceiving the game world through the eyes of the main protagonist in an interesting way for some points of dialogue. It is established early in the game that the player avatar is wearing a helmet with some kind of internal communications system, and the multichannel rendering is used to supplement this idea. When receiving dialogue in this way, the

FIGURE 4.5: A screen-shot from the video game *Doom* (2016) [179] to illustrate what is seen by the player in a first-person viewpoint games.

recorded voice is output from all the loudspeakers in the 7.1 arrangement, other than the centre channel. This helps to create an 'in-the-helmet' effect for dialogue where the speaker is not present, whilst the centre channel is used when the speaker is in the same area as the player avatar to accentuate the idea that they are more directly engaging the player.

The term third-person refers to a POV from which the player sees the in-game avatar they are controlling, usually from behind. For context, game examples include *Super Mario 64* [181], the original *Tomb Raider* [182] and *Gears of War* [183]. There are generally two methods in which game audio is rendered in third-person view games. The first method is similar to what is done in the majority of first-person games in that audio is panned relative to the view of the player. Conceptually, this can be thought of as a camera with an attached microphone, following the in-game avatar through the virtual game world. This is the case in *Journey* [184], where sounds emitting from in-game objects positioned behind the viewpoint (i.e. off screen) are played from the rear and surround channels of the 7.1 arrangement, and the sounds of on-screen objects are output more from the front, in a similar manner to surround-sound mixing for film.

The second third-person rendering method differs to cinema in that the perspective is taken from the centre point of the screen used for visual feedback, rather than the player

FIGURE 4.6: A screen-shot from the video game *The Witcher 3: Wild Hunt* [185] to illustrate what is seen by the player in third-person viewpoint games.

viewpoint. In effect, this gives the impression that sounds within the game world are heard from the POV of the controllable avatar, rather than that of the camera. This means sound emitting game objects positioned behind the avatar, but that are still on-screen, are output from the rear surround loudspeakers, and sounds in-front of the avatar are output from the front left, right and centre channels. This is the case in games such as *Middle Earth: Shadow of Mordor* [186] and *Ratchet and Clank (2016)* [187]. However, it is important to note that the player is able to change the direction in which the avatar is facing, and this has no impact on the panning [2]. Therefore, although the method gives the impression that the game world is heard through the ears of the avatar, it is in-fact heard from a position slightly in-front of the player view, which in most cases is located in the centre of the screen used for visual rendering. The method makes sense in situations where it is desirable for the player to fully immerse themselves in the character of the avatar they are controlling, since sounds from the game world will be heard as if they were placed within it. However, in the experience of the author it can be slightly jarring when NPCs or objects that can be seen on-screen (i.e. physically in-front of the player) are heard from the rear and surround channels.

Nier: Automata [188] is an interesting example of how perspective is used in gaming from

[2]It maybe the case that there are some third-person games where changing the direction in which the avatar is facing does influence the panning, however no examples were found in the research towards this thesis.

both an audio and visual stand-point. The majority of the game is played from a third-person perspective, behind the player avatar, however there are segments within each level in which the perspective shifts to a top-down view and a side-on view. From a top-down view the player sees the game world from above, whilst from a side-view the world is seen on a horizontal plane. In these situations, control within the environment is often limited to only two dimensions (left/right or up/down). When in top-down mode, the avatar acts as the audio listener, therefore sounds pan around the surround-sound system relative to the on-screen position of the avatar. The player is able to move the avatar around the game world, and the relative panning is retained, rather then just being from a central point on the screen. The rendering changes again in the side-view mode in which all sources positioned to the right of the player avatar are output from all right-hand loudspeakers, the same is done to the left. This essentially changes the rendering to stereo output from all the loudspeakers of a 7.1 arrangement. These transitions happen fluidly during gameplay, without breaking engagement.

4.2.3 Use of Centre Channel

As discussed in Chapter 3, for film, the centre channel of a surround-sound loudspeaker array (placed in-between a left/right stereo pair) is, in most cases, reserved as a dialogue channel. This is a method used as a means to give the viewer the impression that dialogue is emitting from the mouth of a speaker, and also to add clarity by keeping speech separate from other sound effects. The use of the centre channel in the context of video games is, however, one of the more inconsistent aspects of multichannel game audio in the way it is used between different titles. This section will introduce and discuss some of the different ways in which the centre channel is used for gaming.

In some cases the centre channel is used in a similar way to film in that it is only used for dialogue for the majority of the time. *Grand Theft Auto V* [166] and *Middle Earth: Shadow of Mordor* [186] are two video game examples that use the centre channel in this way. For these games, the voice of the main player character is always output from the centre channel during gameplay, as well as any other dialogue provided by NPCs that is either used to progress the narrative or provide the player with important instructions, such as through guiding the player towards objectives. All other, unimportant, NPC dialogue is output from other channels, usually depending on the the relative position to the player

avatar and/or audio listener. This is most likely done so that the player does not miss any critical dialogue as a result of it being masked by other in-game sound effects or music. The separation of dialogue is an inherently useful feature in these game examples, and would not be present if the player were to be listening over, for example, stereo, providing a clear advantage in the use of surround-sound for game audio playback. An interesting addition to this concept in *Grand Theft Auto V* is through the use of a small loudspeaker located inside the PS4 gamepad used to control the game. This is used to output telephone calls received by the main character, again potentially as a means to separate dialogue from the core game soundtrack. The technique was also implemented in many titles that made use of the Wii remote controller released for the *Nintendo Wii* console [189].

In other cases, the dialogue might still be output from the centre channel, but other, non-dialogue, sounds are also played. This is the case in *Rise of The Tomb Raider* [190] where important dialogue is mostly output centrally, with some panning to the left and right to accentuate the speaker's in-game position, along with ambient environmental effects. These effects are also output from the other surround channels but at a higher relative volume to that of the centre. Because of this, dialogue is not lost under the ambiences, and the centre channel is therefore in use for the majority of the gameplay. *The Last Guardian* [167], along with the remastered version of *Shadow of the Colossus* [191], are both examples of games that treat the centre channel in the same way as the left/right and surround channels for almost all in-game sound effects. This means that sounds are panned to the centre channel in the same way as they would be to other channels, relative to the in-game position of the emitting object, increasing the number of channels that can be used for sound source placement. The centre channel is also notably used in this way in games that support surround-sound but do not have any dialogue, like *Flower* [192].

A more unconventional use of the centre channel is apparent in the gameplay segments of *Nier: Automata* [188] in which there is dialogue, but it is only ever split between the front left and right channels. The centre channel is instead only ever used for dialogue in cut-scenes and the main character's Foley sound effects, such as impact sounds and grunts when the character comes into contact with an unfriendly NPC. Compared to other examples, this is a strange use of the available channels of a surround-sound system in that the developers intend sound to be output centrally but only in very

specific circumstances. However, part of the game involves switching between different camera view perspectives on the fly (see Section 4.2.2), therefore the decision to only use the centre channel minimally may have been a sensible design choice. This is also the case in *Yooka Laylee* [193] and *Crash Bandicoot N. Sane Trilogy* [194], where the centre channel is very rarely used in comparison to the others available. These two examples are considered to be aimed towards a more 'casual' game-playing demographic and will usually incorporate simple rules and controls, in order for it to be played by as wide an audience as possible.

4.2.4 Music Panning

Music is considered to be an integral part of almost any gameplay experience, for the ways in which it can elicit emotional responses from listeners, thereby supporting the narrative, as well as the ways it can be used as a means to represent different game states [4]. For example, a player navigating their in-game character down a fairly quiet and unassuming corridor, might suddenly find an unfriendly character jumping out from behind a corner, and dynamic changes in music can be used to help enforce these two contrasting states of gameplay. A classic example of dynamic music is in the arcade game *Space Invaders* [5], where the tempo of the soundtrack gradually rises in-sync with game progression, creating an increasingly tense and hurried atmosphere as the player approaches the end-level state. Game music has continued to evolve with the increased implementation of surround-sound formats, and there are generally two methods for playback in modern titles.

The first is similar to what is done in the majority of multichannel film soundtracks whereby a standard left/right stereo mix of the game music is output from the frontally positioned stereo pair. There are many examples of games in which this is the case, but no obvious pattern as to the type (genre) of game, or within titles released by the same developers. *Ratchet and Clank (2016)* [187], *Attack on Titan: Wings of Freedom* [195] and *Dear Esther: Landmark Edition* [196] all employ stereo playback for in-game music and are all relatively different in terms of gameplay and narrative. One could speculate that due to the relative importance of music in games, it is perhaps the case that stereo mixes are used for compatibility reasons so that the majority of players will receive a similar experience between, for example, television loudspeakers, headphones and dedicated

loudspeakers. Of course the majority of popular music is mixed, mastered and released for listening over a stereo system which may be an influence on the way in which game music is recorded, and played back.

The second method uses the full extent of the surround-sound system for music playback. Again, the use of the surround-sound system varies between games and developers. The musical soundtrack from *Middle Earth: Shadow of Mordor* is one of the many examples to output a regular stereo mix over all the loudspeakers of a surround-sound system such that the left channel is output from all those positioned to the left of the listener, and the right channel is output to all those positioned to the right. The centre channel is a sum of the left and right, but in some cases it is not used for music at all whilst music is output from the remaining channels, as is the case in *DOOM (2016)* [180]. In these examples, the relative volume of the music in the surround channels is lower than that of the front stereo pair, which helps to keep the focus of the player forward (in the direction of the visuals) whilst also providing an overall more enveloping listening experience. This heightened sense of envelopment provided by the game music serves to add an element of excitement and immersion to the gameplay experience, whilst not overly distracting from other in-game sound effects. *Horizon Zero Dawn* [197] has a combination of the two methods, with surround-sound music playback in the cut-scenes, and a regular stereo mix played during gameplay. In general, the narrative portions of the game are revealed during these cut-scenes, and the enveloping nature of the surround-sound music mix helps to give these moments an increased sense of importance. Folding the music back down to stereo during gameplay then gives the player a better feel for the in-game world, where mostly ambient and environmental sounds are output from the surround channels, rather than music.

4.3 Video Game Content Case Studies

The listening tests introduced in Chapters 6 and 7 investigate how the use of surround-sound might influence a player's perception of the overall game session. It is therefore necessary to choose video game content appropriate for use as experimental stimuli in a listening test environment. This section first gives a criteria used to choose a selection of games and subsequently provides reviews of four video games highlighting

the suitability for user testing and how the audio implementation might serve to influence player experience. This is done according to the appropriateness of the in-game audio and gameplay elements. The game needs to demonstrate effective use of audio, especially in terms of the spatial qualities and multichannel rendering capabilities. The audio is only deemed to be 'spatially effective' if examples of all the spatial attributes given in Section 3.6 can be identified. Only games capable of 7.1 surround-sound output are considered for this reason. A number of gameplay considerations are also taken into account to ensure potential participants are able to interact with the content, and that similar gameplay experiences might be had between different participants on multiple run-throughs. This will improve the repeatability of any test. To make sure the tests are balanced, it is important that participants play the same section and that in-game objectives are fairly self explanatory and linear (i.e. there is an obvious path and consistent in-game path to follow). All of the following examples were played for testing and review using either a PlayStation 4 (PS4) or PlayStation 3 (PS3) console with 7.1 surround-sound played back over a physical loudspeaker array of Genelec 8040s.

4.3.1 Game Selection Criteria

The criteria were split into two main classifications: audio and gameplay. The audio criteria were used to consider how the specific aspects of the game soundtrack might serve to influence the game experience. The gameplay criteria were needed to ensure that the game would be playable in the context of a listening test.

The audio criteria were as follows:

- In-game examples of the spatial audio attributes identified in Chapter 3, such as sound source localisation.
- 7.1 surround-sound compatibility.
- Third-party acclaim for use of audio from, for example, online reviews.

The gameplay criteria were as follows:

- Repeatability in that the multiple players should receive a similar experience.

- A limited number of 'fail-states', meaning it should be difficult for a player to fail as a result of in-game character death or through not completing an objective.
- The ability to easily restart with little or no back-tracking.
- A simple control-scheme.
- An easy to follow, preferably linear in-game path.

It is important to note that not all the games mentioned over the next few pages adhere to all the points in the set criteria. However they do apply some of the criteria in interesting ways, hence their inclusion.

4.3.2 The Last of us: Remastered

The Last of Us: Remastered [198] is a reworking of the original PS3 game of the same title, developed by Naughty Dog and published by Sony Computer Entertainment, released for the PS4. The game has won a multitude of awards [199] including those relating specifically to audio such as the BAFTA for Audio Achievement [200] and G.A.N.G (Game Audio Network Guild) awards for Audio of the Year, Sound Design of the Year and Best Audio Mix in 2014 [201], giving rise to its acclaim in both the game and audio industry. It has also received acclaim outside of the game industry for its maturity in story telling. The narrative of the game revolves around a fictional post-apocalyptic America where an unknown virus is responsible for either killing most of the population or turning its victims into 'zombie-like' creatures, with remaining cities being occupied and defended by a small handful of the surviving populace. For most of the game the player controls Joel, a middle-aged man grieving the death of his young daughter Sarah, who is tasked with transporting Ellie, a teenage girl with immunity to the virus, across the USA in the hopes of developing a cure.

Audio is rendered in up to 7.1 surround-sound and is compatible with the DTS Digital Surround and Dolby Digital codecs. The player is also able to customise the audio settings in-game, such that the angles used to calculate the amplitude panning between surround channels, match the angles of the player's physical loudspeakers. The feature is rarely seen in games and is done in this case in an attempt to optimise sound source localisation between different living spaces. The game is played from a third-person, over the shoulder perspective, and all sounds are panned relative to the position of the

FIGURE 4.7: A screen-shot taken from *The Last of Us: Remastered*.

player view/camera. The use of surround-sound helps to enrich the overall aesthetic of the virtual environments and settings presented to the player, whilst also being essential to the gameplay. The game encourages the use of audio cues to make certain in-game tasks easier. An example of this is with the main enemy NPCs, referred to as 'clickers' whom the player must either avoid or fight for the majority of the game. In general, If the player is ever caught by one of these clickers, then the game is over. As the name might suggest, these NPCs emit a unique clicking sound whenever they are in the vicinity of the player avatar. This is a useful game mechanic, as the player will often hear this sound before actually receiving any visual feedback. Listening to the game over a 7.1 surround-sound system further enhances this advantage by giving players an idea as to the spatial position of the clicker in the game world, providing more time to prepare for potentially game ending events. With regards to the rest of the game's sound design, a very minimal approach has been taken in a similar way to some horror films, where acoustic space is given to reflect the emptiness of the presented virtual environments, which enhances the tension caused whenever a clicking sound is suddenly heard.

In terms of gameplay, the player usually has to follow well marked paths towards the next objective, with obstacles in between. There are also some exploratory aspects, although it is difficult for the player to diverge from the given path too much, meaning game sessions between different players would be relatively similar. During sections in

which the player is required to fight enemy NPCs, the control scheme can become rather difficult, in that there are a number of different button combinations to be learned in order for the player to perform certain tasks. This would make such sections potentially inappropriate for user testing, in that a less experienced player might find these sections more difficult. The introductory sequence of *The Last of Us: Remastered* is however more appropriate for the purposes of experimentation. The player has only one objective, which is to escape a town in which a disaster is happening by following visual prompts as to which buttons on the control pad need to pressed. The player is required to follow a fairly linear path, making it hard for potentially inexperienced participants to get lost, and the majority of the audio events are scripted and will not trigger until the player encounters a particular section. This ensures similar auditory experiences are had on multiple playthroughs, the only potential difference being the amount of time it takes to navigate to the next area. Furthermore, the number of fail-states in the sequence is considerably low; where even if the player does fail they are able to quickly continue the playthrough with minimal consequence.

4.3.3 Alien: Isolation

Alien: Isolation [202] , developed by Creative Assembly and produced by SEGA, is a first-person survival horror game based on the Alien film franchise. The game takes place 15 years after the events of the original film with the player controlling Amanda Ripley, exploring a decrepit space station. The player eventually learns that the station is threatened by an Alien creature (or *Xenomorph*), hunting down and killing any and all survivors. The player must navigate Amanda through the station, in an attempt to escape, whilst avoiding the Alien and other threats (malfunctioning androids and hostile survivors) with limited supplies. It is not possible to kill the Alien, and limited supplies make the elimination of other threats difficult, therefore the most effective way to play is by avoiding conflict and making use of various hiding places. The tension in the game is greatly enhanced by the use of audio, in much the same way as the original film, which at the time was celebrated for its sound design and musical score [203]. The developers intentionally made use of lo-fi sound effects inspired by the original film, to the extent where effects from the film made it into the final game [204]. This decision, along with professional voice acting for in-game characters, gives the game a realistic

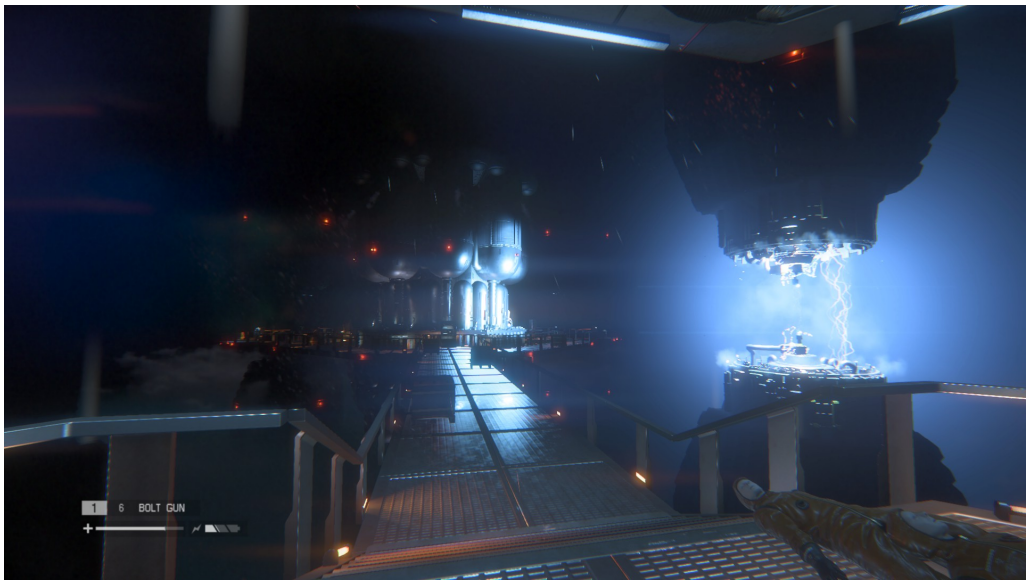


FIGURE 4.8: A screen-shot taken from *Alien: Isolation* showing the sound emitting towers in the main reactor room.

and authentic feel, creating an immersive game experience.

The game outputs up to 7.1 surround-sound using a Dolby Digital codec and is played from a first-person viewpoint. Sound sources are placed accurately and appropriately in the virtual space and react as expected in relation to the movements of the player avatar. Because of this, when playing in surround-sound, the player is able to estimate the spatial location, and distance, of the *Xenomorph* (or any other potential threats) far more easily than when playing the game using a mono or stereo system, based on the localisation between surround-sound channels. This allows the player to better plan diversions and their escape route, thus providing an advantage for navigating towards objectives unhindered. The game is a good example in which there is a potential for player performance to be improved as a result of the surround-sound mix. This also adds a sense of realism to the environments presented to the player. A sense of envelopment is created by the simulated acoustics of the virtual environment when listening over a surround-sound arrangement, suitably establishing feelings of tension by conveying the idea that the player is exploring a vast and lifeless structure. Throughout most of the game, the only sounds that will be heard are the footsteps of the player avatar, and the resultant acoustic reflections from around the environment. This is a common audio trope found in the horror genre, where extended segments of silence (or minimal use of audio) are often employed in order to make the 'scarier' sounds

FIGURE 4.9: A screen-shot taken from *Alien: Isolation* illustrating the radar held by the protagonist. Sounds from this radar can be heard from a loudspeaker located in the gamepad held by the player.

stand out, and therefore force an emotional response from the player [205]. A particularly noteworthy environment is at a later stage in the game where the player finds themselves in what is described as the 'main reactor room', a huge cavernous space surrounded by giant towers emitting electricity. These towers are spaced at various points around the player avatar and continually emit loud, explosive sounds, which, when heard simultaneously and combined with the in-game positions, give a real sense of depth and width to the environment (See Figure 4.8).

The gamepad used to control the game also provides audio feedback through a small internal loudspeaker, providing an extra channel to the surround-sound. During gameplay a hand-held radar-like device (see Figure 4.9) is used by the player avatar in order to track the movements of potentially dangerous NPCs. The loudspeaker in the gamepad emits short beeps when enemies are nearby, simulating the function of the in-game tracker, and giving the player the impression that they are holding a physical piece of equipment from the game. The player is also able to plug a camera and microphone into the PS4 console to supplement this feature. Real-world movement and sound from the player are then detected and fed back in to the game as a way to influence decisions made by the artificial intelligence (AI) controlling the *Xenomorph's* actions. Although this doesn't relate directly to the use of surround-sound for the game

audio, it is interesting to see the other ways in which developers try to make players feel a part of the game world through the use of audio content.

Even with the many positives concerning the game's use of surround-sound, and audio design in general, the gameplay is potentially not ideal for the purposes of experimentation, especially in the case of participants who might be lacking in game playing experience. Avoiding the game's threats is a challenge, even on lower difficulty settings, and getting caught often results in a fair amount of back tracking, especially if the player does not regularly save their progress at specified points. It is therefore difficult to guarantee different participants, on multiple playthroughs, will have similar audio experiences. The game does feature a more repetitive 'challenge mode' in which the player is required to complete a set of tasks whilst being hunted by the *Xenomorph* in an allocated time limit, however, the game's complex control scheme is still not ideal. The game would be appropriate as a test environment if it was ensured that participants were well experienced in the game itself, based on past experiences, or through a number of training sessions.

4.3.4 P.T.

P.T. [206] (an abbreviation of *Playable Teaser*) is an experimental demonstration made for the PS4 as an interactive teaser for a game that is no longer in development. *P.T.* was developed by Kojima Productions and published by Konami with involvement from esteemed game designer Hideo Kojima and film director Guillermo del Toro, who say that the game was intended as a short interactive 'film-like' experience in the style of an independent developer [207]. Like *Alien: Isolation*, it is a first-person horror experience in which the player controls an unknown male protagonist trapped inside an infinitely looping L-shaped corridor. The objective is to try and escape this corridor, by continually navigating towards a closed door and its end. Upon each successive loop, the narrative gradually begins to unfold, revealing that the previous owner of the house murdered his family who are now ghosts. To begin the next corridor loop the player usually has to solve some sort of cryptic puzzle, whilst increasingly unsettling things happen around them. It is nearly impossible to complete the experience without following a specific set of instructions.

Up to 7.1 surround-sound is supported, but it is not clear what (if any) codecs are used. Sound source localisation is used to great effect, in that the placement of realistic effects, such as a door creaks and the humming of a wall mounted light fitting, help to place the player in the midst of the environment. A vintage style broadcast emitting from a radio in the corridor is used to give the player information regarding the narrative. The broadcast becomes clearer as the player navigates their avatar towards the source, showcasing a good use of distance simulation. The effect is suitably unsettling within the context of the game, as it forces the player to continue moving through the corridor, towards the source. This is a good example of spatial audio effects being used to progress game narrative. As the game continues into a more psychologically frightening experience, the use of surround-sound begins to reflect this. An early example is when the player navigates past a bathroom door and loud banging sounds are output from the side and rear surround channels. This is the first deliberate 'scare' in the game, after which the soundscape becomes increasingly eerie. This could be thought of as a reflection of the main character's mental state, as after this first frightening experience, it becomes unclear whether some sounds are actually emitting from the corridor or are just in the character's head. Examples include the footsteps and breathing of an unknown character behind the player and the muffled sounds of a disembodied crying baby. Often these sound cues are used to attract the player's attention to some clue that will allow them to progress to the next stage, again showcasing how surround-sound might be used as a narrative device.

FIGURE 4.10: A screen-shot taken from *P.T.*.

Considering the simplicity of the game's environment (a repeating L-shaped corridor), *P.T.* would be a good choice of content for experimentation, even for inexperienced video game players. It is unlikely for participants to lose their way, and there is a high chance similar experiences will be had on multiple playthroughs. This is especially the case for approximately the first 10 minutes of gameplay, where there are not many puzzles to be solved. The control scheme is fairly simple and similar to a variety of other first-person perspective games, where the left analogue stick on the gamepad is used for navigation and the right for rotating the player viewpoint. The 'X' button on the gamepad is also used for simple interactions with in-game objects. A set of instructions guiding players through the corridor and giving puzzle solutions, with specified button prompts, would also help to streamline the process of testing. A disadvantage with *P.T.* is that further into the game the player is required to perform extremely precise actions, such as walking an exact number of steps after hearing a particular sound cue and whispering into a microphone attached to the PS4 gamepad. Therefore *P.T.* would be of most use in potentially shorter tests, where it is unlikely for participants to reach such a stage. The genre also needs to be taken into consideration in that some participants might find the experience to be rather unsettling, due to some of the audiovisual elements and subject matter. It would therefore have to be made clear to participants as to the graphic nature of the game, before considering their involvement.

4.3.5 Ratchet and Clank: Tools of Destruction

Ratchet and Clank: Tools of Destruction is a third person action and 3D platformer game developed for the PS3 by Insomniac Games and published by Sony Computer Entertainment, released in 2007. Games of this genre usually involve exploring a 3D environment filled with NPCs controlled by the AI, and some element of jumping between platforms to gain access to new areas, hence the term '3D platformer'. The game contrasts greatly with those previously discussed, having stylised cartoon visuals and an over-the-top, tongue-in-cheek narrative. The game was chosen partly for this reason, to investigate the use of surround-sound in a game aimed at a 'casual' audience. The player controls Ratchet and Clank, two cartoon characters resembling a Lynx and robot respectively. The game is split into separate missions which usually involve following a linear path to get from point A to B whilst defeating waves of game

controlled NPCs, and picking up in-game collectible items along the way. Every so often the player is required to complete a 'boss' fight, as is the case with most other games of a similar genre. The game is played from a third-person viewpoint, behind the player avatar, and outputs up to 7.1 surround-sound using the Dolby Digital codec. Sounds are panned relative to the centre of the screen, to give the impression that it is heard from the position of Ratchet and Clank.

A hyper-realistic approach has been taken in regard to the sound design, meaning that sounds effects occur when one would expect bu they have been exaggerated in some way, which keeps in theme with the in-game universe and colourful visual style. The surround-sound mix in *Ratchet and Clank: Tools of Destruction* enhances the spatial definition of separate sound effects, in what is often a relatively busy soundscape. Each in-game object the player can interact with is attached to at least one sound effect, which combined with combat and environmental/background effects can result in a sometimes overly fatiguing and uncomfortable experience when listening to the game in stereo or mono. The separation provided by the mapping of sound effects to the seven loudspeakers of the surround-sound arrangement help to define different effects and place the player in the centre of the action. The first level in the game is a good example of such surround-sound implementation. The environment is a futuristic city located on a fictional planet, containing such sound effects as hover cars, sprinkling

FIGURE 4.11: A screen-shot from *Ratchet and Clank | : Tools of Destruction*, sourced from [208].

water features and hissing pipes. At the beginning of the level, enemy NPCs approach Ratchet and Clank from multiple directions, which is reflected in the placement of their related sound effects over the surround-sound system. Later in the mission the player controls Ratchet and Clank along a train track whilst vehicles rush towards them. The sounds of the vehicle can be heard approaching the camera and eventually flying past the view, giving the impression of depth and width.

Due to the game's target audience and Pan European Game Information (PEGI) rating (7+) most elements of the gameplay are very easy to grasp making it perfect for user testing, especially if younger participants are to be considered. If the player does fail an objective, they are able to try again relatively quickly, minimizing back tracking and repetition. The control scheme is fairly simple and should be familiar to participants who may have played similar third-person content of the same genre. The first mission, as outlined previously, lasts for around 10-15 minutes and is relatively straightforward in terms of the path the player is required to follow. The actions that must be performed to achieve certain objectives are also relatively self-explanatory.

4.4 Summary

This chapter has discussed the different ways in which multichannel audio is implemented in video games, beginning with a history based on commercial video game systems and then investigating some specific video game examples. 5.1 and 7.1 surround-sound formats are widely used in gaming and this is currently being expanded with the introduction of Dolby Atmos in a handful of titles. Binaural synthesis and VHT systems are less common, however, their use in VR applications is becoming more prominent. An interesting point to take from Section 4.2 is that there are many ways in which surround-sound can be implemented in gaming, sometimes adhering to conventions established in the film industry but at other times doing something completely different. For example, *Middle Earth: Shadow of Mordor* is similar to film with regards to dialogue being output mostly from the centre channel, however, the music is output over the entire 7.1 arrangement, which would be unusual in film. This suggests that, as yet, there are no established standards in how surround-sound is used in gaming. A lack of common conventions may perhaps arise because different

types of games require different types of interaction from the player, whereas in most films an audience is expected to participate in a similar manner upon different viewings. Standard surround-sound practices may therefore lend themselves better toward film applications.

Content intended to be used as experimental stimuli in the later parts of the thesis has also been short-listed in the form of four case studies focusing on commercially available titles: *The Last of Us: Remastered*, *Alien: Isolation*, *P.T* and *Ratchet and Clank: Tools of Destruction*. These games offer some interesting examples in terms of sound design, especially in regard to the use of surround-sound. Gameplay characteristics were also considered to ensure that these titles are suitable for user testing in terms of the ease of control, the objective of the player and difficulty levels. An interesting point to take from these case studies is that the titles considered predominantly fall somewhere within the action, survival and horror video game genres. These types of games form a relatively large part of the overall gaming market, implying that their development might receive higher budgets. This would therefore allow a more significant part of the game making process to be dedicated to audio resources.

Chapter 5

Determining the Player Experience

This thesis has, so far, presented examples of multichannel audio systems, both physical and virtual, and how the spatial characteristics of audio rendered over them can be effectively conveyed. Also, there has been discussion considering the way in which multichannel audio might be used in the context of video game virtual environments and gameplay, and how this can be different depending on the style of game. However, there has not yet been any consideration as to how these multichannel systems might actually influence a players judgment concerning the quality of their experience, whilst engaged in playing the game. The purpose of this chapter is the introduce the concept of *Quality of Experience* (QoE), a phrase often associated with a user's judgment of a piece of multimedia content, and how this might relate to multichannel video game audio. This will help to form the foundation for the perceptual listening tests presented in the later parts of this thesis.

After defining QoE, two different measurement methodologies are introduced as well as a workflow/framework designed to be used as an aid in designing QoE tests. The idea of using preference as a metric for determining overall QoE is introduced, and this will be returned to later in the thesis. This framework provides a groundwork for determining an experimental process by ensuring that appropriate stimuli are chosen and suitable metrics are used to gather subject/user responses. The definition for QoE extends to many applications and types of multimedia content beyond audio and video games, but this chapter will focus more specifically on how characteristics relating to audio, such as the spatial attributes that can be conveyed over a multichannel listening system, are considered to be an influencing factor. Examples from the literature with respect to audio

quality and spatial audio quality influencing user experiences are given.

5.1 Quality of Experience

An individual's Quality of Experience (QoE) is considered when it is desirable to know the extent to which an engineered product/system is liked. The core idea underpinning studies into QoE is that people are able to form subjective quality judgements through experiencing the content of interest. Elements of the experience can then be altered between exposures, and subject responses retaken, as a way to identify those characteristics of the experience that have an influence on the individual's overall quality judgement. In playing a video game, a player experiences all the elements used to construct the game as a single entity, including (but not limited to) the graphical fidelity, narrative, user interface, gameplay mechanics, interactive systems, music and audio effects. It is not unreasonable to think that creators of video game content will hold the QoE of potential players in high regard, since, in general, enjoyable game experiences are desirable. By studying QoE, the more influential underlying elements of the gameplay can be identified, scrutinised and changed in ways that might more positively influence the experience of the player. This chapter has a particular focus on the ways in which audio is an influencing factor on user experiences, and how it can be manipulated to influence this.

5.1.1 QoE Definition

Quality is defined by Jekosch [209] as '*the judgement of the perceived composition of an entity with respect to its desired composition*'. Used in the context of experiencing a product/system, this quality judgement can be understood as a subjective evaluation of the stimuli's combined percepts, i.e. the experience. Raake and Egger [210] expand on this by giving a full definition[1] for Quality of Experience, as follows:

> *The degree of delight or annoyance of a person whose experiencing involves an application, service or system. It results from the person's evaluation of the fulfilment of his or her expectations and needs with respect to the utility and/or enjoyment in light of the person's context, personality and current state.*

[1] It is important to note that the definition is derived from previous definitions/standards. A detailed explanation as to how these authors arrived at this current definition is given in [210].

Here, a *person* refers to any individual who is exposed to the stimuli, either directly or as a spectator/onlooker, and their *experiencing* is driven by those characteristics of the stimuli (the *application, service or system*) that are perceived. The term *utility* is used in reference to Kahneman's concept of *experienced utility* [211] which proposes the theory that individuals naturally evaluate a given experience on a scale of good to bad based upon whether it was or was not enjoyable. It is important to note that with this current definition the experience might also involve both circumstantial and contextual elements that can influence the overall quality judgement. For this reason it is difficult to measure pure QoE for a specific experimental stimuli, and this observation will be covered more in Section 5.2.

5.1.2 Utilitarian QoE assessment

Along with the definition of QoE, Raake and Egger propose two distinct types of quality assessment [210]. These are based on the natural ways in which an individual might react to some experimental stimuli. The first is a *utilitarian quality assessment*, which also relates to Kahneman's theory of *utility*, as used in the definition for QoE. The aim is to gain an overall impression of the perceived quality of the stimuli under investigation. The elements that make up the experience are treated as a singular entity, rather than as several different percepts or characteristics, and in this way the experience as a whole can be more easily assessed with a simple 'good' or 'bad' rating. Measuring the experience in this way has its benefits in that it is a relatively easy test to administer given that the concept of 'good' or 'bad' will be understandable over a wide range of individuals/demographics. It is also easy to see how this might be appropriate for experiments that may require participants to play a video game for a substantial amount of time, which is the case in Chapters 6 and 7. By asking for an overall judgement after the game session, participants can more easily focus on the game task without risk/fear of interruption.

As a utilitarian method, preference tests provide a close approximation to overall QoE as they take into account the presentation of the stimuli in its entirety, rather than focusing on specific characteristics/aspects [212–214]. This assumes that if one experimental condition is preferred over another, then a more fulfilling QoE is had. This isn't to say the the experience is bad in the less well regarded exposure, only that the

other has been improved in a noticeably positive way. Work by Choisel and Wickelmaier [126], [215] implies that there is a relationship between the degree to which a multichannel listening system is preferred and how well it can convey auditory sensations to a listener. This therefore suggests that a listening condition that is perceived to have a high spatial sound quality will be preferred over one that is not.

Preference tests between two different experimental stimuli can be easily administered using a paired comparison design [216], after the individual is exposed to both experimental conditions. The preference for one condition over another (in this example A compared to B) is rated using a 7-point paired comparison scale. For each comparison the scale is structured as follows:

(3) Strong preference for A

(2) Preference for A

(1) Slight preference for A

(0) No preference

(1) Slight preference for B

(2) Preference for B

(3) Strong preference for B

This paired comparison design assumes that the preference rating given for one condition will yield the opposite rating for the other condition [217]. For example, if the participant feels a strong preference towards stimuli A, it is assumed that this means there was a strong non-preference for stimuli B. Numerically, this would translate as a preference score of 3 for stimuli A and a preference score of -3 for stimuli B. If neither condition is preferred then participants have the option to choose 'No preference', giving scores of 0 for both conditions. The design in useful as it only requires the subject to give one rating, after they have been exposed to the stimuli using both listening conditions.

The draw-back in using only utilitarian/preference assessments is that this does not provide any information as to why the preference is given. For this reason QoE assessments also take into consideration the individual characteristics of the stimuli in the form of analytic assessments.

5.1.3 Analytic QoE assessment

The second type of assessment proposed by Raake and Egger [210] is an *analytic quality assessment*. Whereas utilitarian methods could be considered as high-level assessments, analytic methods are based around an individual judging the quality of an experience through pre-determined low-level characteristics, or influencing factors (see Section 5.2). The assessment is decomposed into individual characteristics/attributes that have been hypothesised to influence the experience of the stimuli in some way. These are then rated individually on separate scales, with the idea being to identify those attributes that contribute significantly to the overall (utilitarian) quality judgment.

Although there are no specific examples of this kind of assessment relating directly to audio in the literature, the theory can easily be applied to existing audio rating methods. The rating scale for *basic audio quality* given in the ITU recommendation BS. 1284 is an analytic assessment, as it is suggested that audio content should be rated according to individual attributes, chosen by the investigator [218]. Participants are asked to assess each chosen auditory attribute (such as the spatial attributes defined in Section 3.6) according to a five-point scale, either during or after exposure to the experimental stimuli. The method offers a simple approach to analytic audio assessment, as all attributes are rated on the same scale, increasing comprehension between participants. The proposed scale is as follows:

(1) Bad

(2) Poor

(3) Fair

(4) Good

(5) Excellent

More complex examples of audio assessment that could be related to analytic quality assessment include the MUltiple Stimuli with Hidden Reference and Anchor (MUSH-RA) test [219] and the methods proposed in ITU recommendation BS.1116 for small audio impairments [220]. Tests using these methods are usually designed in a way that easily allows participants to switch between experimental stimuli, in order to make comparative ratings. Ideally, this means audio material/stimuli should be relatively

short in order to allow participants to recall differing aspects between exposures. However, the experimental work presented in Chapters 6, 7 and 8 are based around users interacting with video games, causing the length of exposure to any one stimuli to be significant. The relative simplicity of the *basic audio quality* scale makes it more appropriate for the video game stimuli used in these Chapters, in comparison to the other methods mentioned here.

5.1.4 QoE test framework

A framework for subjective QoE assessment is proposed by Agboma and Liotta [221]. Although the framework is presented for mobile streaming systems, it is still useful in that it offers a structured approach for QoE assessment that can be easily generalised for use in other applications like listening tests. Here, the definitions for each step of the framework have been adapted to accommodate for some of the QoE concepts covered so far, with an emphasis on listening tests for game audio.

1. **Characterise the application**: Identify characteristics/attributes that are believed to influence QoE.
2. **Design and define test matrix**: Specify the objective parameters (or independent variables) that can control the behavior of the attributes identified in step 1. These are the parameters that are changed during the test. For example this can be the type of multichannel system used for listening.
3. **Specify test-bed and materials**: Choose/generate experimental stimuli (e.g. audio stimuli or a video game) that is representative of steps 1 and 2.
4. **Carry out subjective assessments**: Data is gathered by asking participants to assess the stimuli according to the utilitarian and analytic methods discussed previously, during or after exposure. It is important that participants are not made aware of the objective parameters being changed.
5. **Analysis of results**: Unreliable results are removed and statistical analysis is performed to find any significant differences in the data. The aim at this stage is to determine whether changing the objective parameters of the stimuli has any perceptual impact on the quality of attributes from step 1.

6. **Statistical modelling techniques**: A predictive model is generated based on correlating the data analysis with the objective parameters. Predictions can then be made as to how users might react to product changes, without the need for further subjective testing.

7. **QoE management strategy**: The results and statistical models are used to influence further decisions in the development of the product.

This framework is used to aid in the development of the experimental methods presented in Chapters 6, 7 and 8, providing a consistent approach across all of these proposed listening tests. This involves switching between objective parameters (in this case of multichannel audio rendering systems) that are expected to change the way in which attributes of some stimuli are conveyed to a user. The subjective assessment (based around combining utilitarian and analytic quality assessments) then makes it possible to firstly come to conclusions as to whether these changes are even perceptible by a user, and secondly, if the quality judgement reflects this. Gathered subject responses might provide some clarity as to whether the self-reported experience of a user can be influenced by simply changing the way in which specific characteristics are communicated. An important point to take from this methodology is that the experimental stimuli should clearly showcase the attributes under consideration, and that these attributes are impacted by changing the objective parameters. In the context of the experimental work presented later in this thesis, this is considered in the content selection of Chapter 6 and the development of a custom, interactive and game-like localisation task in Chapter 8 (Section 8.4.1).

5.2 Influencing Factors on QoE

According to the Qualinet white paper on definitions of QoE [222], any attributes of the stimuli that are thought to significantly impact the overall quality judgment are referred to as *Influencing Factors* (IFs). Thinking in terms of IFs helps to inform steps 1 and 2 of the aforementioned framework, where it is desirable to know what might have a major impact on the QoE outcome. This section will introduce some of the dominant IFs believed to influence the QoE of multimedia applications, specifically focusing on those that are, or might be, related to audio quality. Examples are also given on how

multichannel audio specifically can influence overall audio quality, and how this might impact QoE.

Three top-level IFs are identified in the Qualinet white paper comprising of 'human', 'contextual' and 'system' factors. Human factors are centred around the user's background whilst also considering their physical, mental and emotional disposition, and contextual factors relate to the environment in which the product/stimulus is exhibited. System factors relate to the technical aspects of the experience and the way in which it can be exhibited to a user, through a combination of visuals, audio and interaction (i.e. a multimodal experience). Reiter et. al go on to expand on the idea of 'system' factors by defining further subsets of IFs [223]. There are many IFs in the list and those relating specifically to audio are listed here:

- Content-related: Audio bandwidth, dynamic range
- Media-related: Encoding, sampling rate, synchronisation
- Network-related: Bandwidth, delay, compression
- Device-related: Channel-count, headphones, loudspeakers

These subsets of influencing factors provide an indication that the way in which audio is presented is believed to impact the QoE of multimedia applications. These IFs seem to suggest that objectively degrading or enhancing the audio in some way is most important when considering QoE (this is discussed further in Section 5.2.1). More specifically, considering the subset of 'device-related' IFs (especially in regard to channel-count) then it can be expected that the use of multichannel audio in the context of a game will have an influence on the player's QoE (see Section 5.2.2).

It is important to note that the definition for QoE is designed to encompass a broad range of multimedia content, meaning there are many different types of stimuli that can be covered, making pure QoE difficult to measure in a single, universal rating system. This point is reflected by Raake who suggests adapting the QoE test design based on the content/question under consideration, meaning for an audio based test the QoE can be inferred by thinking in terms of sound quality [224]. This helps to narrow the assessment by focusing on more specific elements of the content and lends itself well to the analytic QoE assessment methodology outlined previously. For this reason there is

some lack in consistency in the types of metric used to actually rate QoE in the literature, mostly due to the diverse application of QoE and types of content. However, there is a general trend in that the QoE is often inferred, or predicted, by analysing subjective assessments of some pre-determined attributes.

5.2.1 Audio as an influence on QoE

Based on the IFs identified previously, audio characteristics can contribute significantly to whether a piece of multimedia is perceived to offer a good or bad experience. The prominent forms of multimedia covered in the literature, in the context of QoE studies, are those relating to mobile device/personal computer applications, such as streaming, and high definition television (HDTV) technologies. Although there is no direct relation to game audio, the same fundamental observations can be easily applied. This section will provide evidence from the literature pertaining to how objective degradations in audio quality can influence QoE judgments.

By degrading the audio track of audiovisual content through band-limiting, Beerands and De Caluwe have shown that audio significantly influences the overall impression of an audiovisual experience, such as a film [225]. Their results suggest that even when the video remains relatively unaltered, the deliberate degradations in audio quality result in a negative opinion of the experience. Davis et. al also provide evidence to suggest that a more enriched soundscape might help to compensate for poor visual quality in an interactive virtual environment experience [226]. In this study participants were less successful at recalling objects in a virtual room when the audio was played back at a lower sample rate and bit depth. It is, however, important to note that there is also research to suggest that well rendered visuals can sometimes compensate for degraded audio. In a study by Rahayu et al. participants were asked to rate visual quality of content in the presence of high and low quality audio and no audio [227]. There was found to be no significant difference in the quality ratings between the three conditions. Both Welch and Warren, and Hollier and Voelcker have also presented similar results in that the visual elements of multimedia content dominate an individual's quality judgment even when the objective audio quality is relatively poor [228, 229].

The synchronicity of audio and visual feedback in audiovisual based content/applications is also considered to be an important audio related influencing

factor on QoE. Asynchrony of audiovisual material occurs when the user begins to notice delays between what is shown to them on screen and any associated audio, for example, when an actor's recorded voice track does not match the movements of their lips. The perceived experience is often judged to become increasingly unsatisfactory as the delays between audio and visual information increase [230], [231]. The same can be said for angular mismatches which occur when the perceived spatial position of a sound source does not match that of the supposed origin (based on visual cues), where displacements of more than 11° become annoying [232]. In addition to this, jitter and lag (the audio noticeably stops and starts beyond the user's control) are considered to be unacceptable in most situations [233, 234]. In cases where annoyances, such as those presented here, persist users are more likely to prematurely terminate an audiovisual experience [235].

5.2.2 Spatial audio as an influence on audio quality

The examples presented above illustrate how the presence of audio might impact a multimedia experience, especially with regard to deliberate audio degradations (through digital processing) and the negative impact that can have. More specifically, it is well established that the spatial attributes of audio are an important factor in its perception, and as discussed in Chapter 3, these can be reliably presented to a listener over a multichannel listening system. This is supported by Letowski who states that 'sound quality extends beyond just timbre', and that spatial audio attributes should be considered when performing subjective audio quality assessment [18]. From this concept, Letowksi presents the MULtilevel auditoRy Assessment Language (MURAL) model to provide a basis for the different auditory attributes that can be used for audio quality assessment. The MURAL model is presented in Figure 5.1, which demonstrates that the number of attributes relating to spaciousness is significant. There is evidence in the literature to back up the MURAL model where experiential work by Rumsey [236] et al. and Zielinksi et al. [237] suggest that spatial quality accounts for a significant percentage of a listener's perception of overall audio quality, when listening over multichannel playback systems.

The concept of spatial attributes impacting perceived audio quality is elaborated further by Le Bagousse et al. through the categorisation of twenty eight commonly used terms

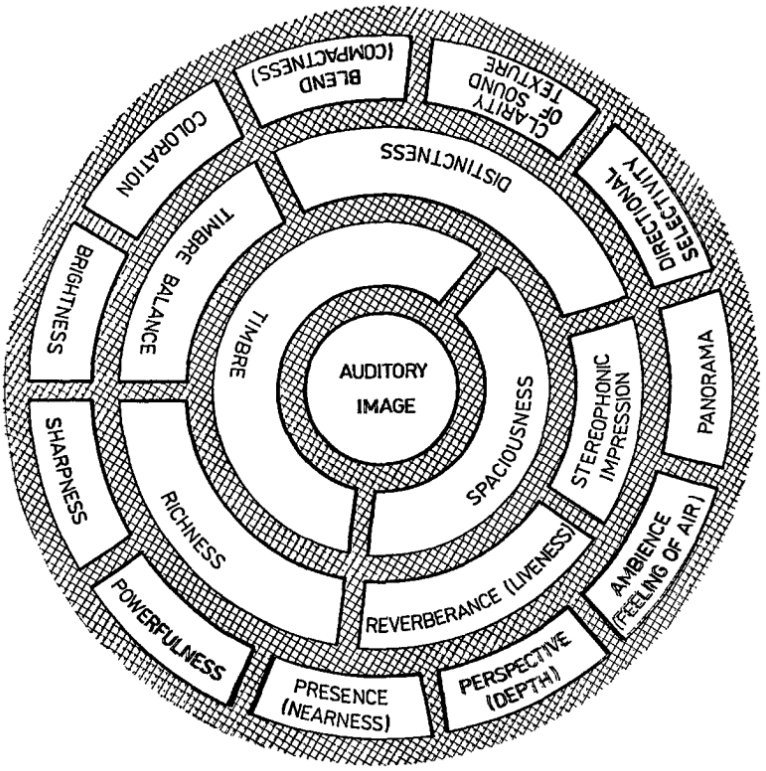


FIGURE 5.1: The MULtilevel auditoRy Assessment Language (MURAL) model, proposed by Letowski, made of of auditory attributes believed to significantly impact the perception of overall audio quality. Spatial (spaciousness) attributes make up a significant portion. Reproduced from [18].

in audio assessment into distinct attribute 'families' [117]. These families are presented in Table 5.1. The work demonstrates agreement with the MURAL model in that two main attribute families, timbral and spatial, are identified. A third family (Defaults) is also identified relating factors external to the sound source, such as the physical space in which the audio is presented. In experimental conditions, the defined attribute families were found to be related to impressions of overall sound quality [118]. The identified auditory attributes form a strong foundation for audio quality assessment using the analytic methodology defined in Section 5.1.3, in that it is generally good practice to rate the spatial quality of audio stimuli using individual attributes, rather than attempting to rate the audio quality as a whole [113].

Defaults	***Timbral***	***Spatial***
Background noise	Fidelity	Reverberation
Noise	Hardness	Spatialisation
Distortion	Richness	Spatial distribution
Disruption	Homogeneity	Localisation
Hiss	Sharpness	Width
Hum	Tone colour	Distance
	Colouration	Envelopment
	Brightness	Depth
	Clarity	Immersion
	Dynamics	
	Realism	
	Stability	

TABLE 5.1: The audio attribute 'families' identified by Le Bagousse et al. intended to be used in audio quality assessment [117].

Loudspeaker Rendering

Work by Berg and Rumsey [19] has considered the use of verbal attributes, such as those discussed in Chapter 3, as a way to define different multichannel audio playback systems. Participants were, in their study, presented with the same audio samples played over mono, stereo and 5.0 loudspeaker systems. The participants' task was to use some pre-defined auditory attributes, identified by the authors in a previous experiment [20], to rate the different stimuli. Through this, it was determined that listeners are able to distinguish the objective spatial differences in recorded audio material, when it is played back over different loudspeaker arrays. The majority of attributes were rated as expected, in that mono was consistently perceived to be worse at conveying those attributes relating to spatiousness, such as *envelopment*, *room size* and *room width*. The outcome of this work suggests it is expected that listening formats with few audio channels, such as mono, are perceived to have low spatial audio quality and that spatial audio attributes are appropriate for determining this.

Dewhirst et. al [21] have proposed an objective measuring system to determine how well different loudspeaker arrays can convey spatial attributes, by computationally modeling the propagation of sound output from different loudspeaker positions. Notably, mono, stereo and 5.0 surround-sound [2] were compared, as well as wave-field synthesis (WFS) and higher-order Ambisonics (HOA). For the three attributes assessed (*localisation*, *width*

[2]The study by Dewhirst et al. [21] uses the term 3/2 stereo, which is equivalent to 5.0 surround-sound, i.e. 5.1 surround-sound but without the LFE channel/subwoofer.

and *envelopment*), the performance of mono was found to be significantly worse than that of stereo or 5.0, although WFS and HOA were generally the best overall. These results also imply that, in general, listening formats with higher channel counts are better at conveying spatial attributes.

Headphone Rendering

For headphone listening, multichannel game audio is very often down-mixed to left/right stereo, but can also be rendered using a VHT system in an attempt to retain the separation between different audio channels (see Chapter 3). A number of studies have however shown that VHT systems are not always considered to be better than stereo down-mixes. A study by Lorho and Zacharov [100] compares a number of virtual 5.1 surround-sound systems with stereo down-mixes of the same audio material. A range of audio stimuli were used, including a video game. However, participants did not directly interact with the material during the test, as would be the case in an interactive game. None of the virtual surround-sound methods were found to out-perform the down-mix, and in some cases the down-mix was slightly preferred.

In a similar study, various virtualisation methods for 5.1 surround-sound broadcast material have been compared by the BBC [238]. This included systems using both individual and generic BRIR measurements (see Chapter 2). Again, the overall sound quality of the processed material was never perceived to be improved in comparison to a stereo down-mix, used as an experimental anchor. Sousa [239] investigated the subjective spatial quality of a 16-channel Ambisonic system rendered binaurally. Although the binaural system was considered to have enhanced spatial quality, a stereo down-mix was still preferable among participants.

Listening tests presented in Chapters 6 and 5 are similarly structured to the previously mentioned studies, in that different loudspeaker systems are compared. The spatial attribute rating task will however be subjective, and simplified because participants are also required to play a game. Having a too complicated audio rating task might distract from the gameplay. Since video games, or participant interaction were not considered in these previous studies, it will be interesting to find whether similar comparisons can be made between the three experimental listening conditions.

5.2.3 Game Audio Quality

The Interface, Effect, Zone and Affect (IEZA) framework is one of the few examples in the literature specifically geared towards improving player experiences through effective audio implementation [240]. The framework is defined by two dimensions; *diegesis* and *expression*, see Figure 5.2. The former concerns whether audio is perceived to be part of the presented game world (diegetic) or not (non-diegetic). Expression refers to whether sound is triggered by the player themselves through interaction (activity) or if it can be heard regardless of player actions (setting). The framework also consists of four categories, superimposed onto the two-dimensional *diegesis*/*expression* space: Zone, Effect, Affect and Interface.

- **Zone** sounds are linked in some way to the game's setting or environment, often in the form of environmental effects and ambiences.
- **Effect** sounds are those associated with specific game objects and sound sources in the game world. Depending on the game genre and aesthetics, these might be realistic or non-realistic (such as synthesised bleeps).
- **Affect** refers to how audio, usually music, is used to express the non-diegetic aspects of the game. In most cases these are intended to evoke an emotional response from the player.
- **Interface** sounds are used to give auditory feedback as a result of interactions separate from the game world. These include actions such as navigating and selecting options from a game menu.

These terms describe the elements in a game (like a playable character or the user-interface) with which specific sounds should be associated. They also provide a check-list of in-game audio systems that developers should take into consideration for a satisfying game experience. The framework is a useful tool for game developers in that it ensures that basic audio feedback is in place, based on what is expected by an end user for a satisfactory experience. The authors of IEZA claim that in employing the framework there are noticeable benefits such as 'richer sound design...better understandable sounds...and more innovative design', thus accounting for high quality game experiences.

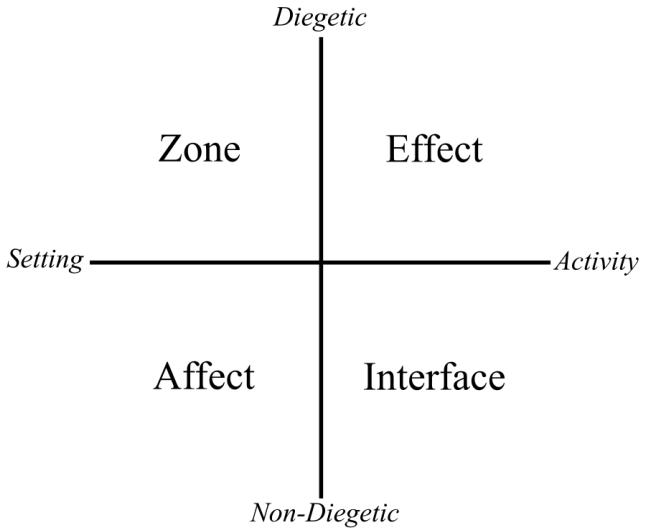


FIGURE 5.2: The *diegesis/expression* space suggested in the IEZA framework for satisfactory game audio design, reproduced from [240].

5.3 Summary

This chapter has introduced the concept of quality of experience (QoE) and how the audio attributes of multimedia content act as influencing factors (IFs). Evidence is also given to suggest that spatial audio attributes can have a significant influence on the perceptual quality of audio. Therefore it could be hypothesised that enhancing game audio by playing it over a multichannel listening system (that is more able to convey spatial attributes) will serve to positively impact the overall game QoE. There is, however, little evidence of this being the case in the literature in the context of video games, and hence part of the motivation for this thesis.

The QoE framework introduced in this chapter will be used to inform the experimental work presented in the next chapters. The two types of quality assessment, *utilitarian* and *analytic*, will be used to form conclusions on how multichannel game audio might impact the player experience. Preference tests will be used as a utilitarian metric to gain an overall impression from the player after a game session. Spatial attribute ratings will also be used as an analytic measure. Those attributes defined in Chapter 3 will break down the sound quality assessment used in the later chapters to infer the specific qualities of a multichannel listening system that might influence the perceived experience.

Chapter 6

Perceived Spatial Quality and Player Preferences

This chapter presents the first of three perceptual listening tests exploring how the characteristics of multichannel game audio might influence video game player QoE. The aim of this chapter is to determine whether a video game session is more preferable if a player believes the system used for audio playback is able to better convey the spatial qualities of the game soundtrack. The study was based on the premise that the spatial characteristics of audio content contribute significantly towards a listener's perception of overall sound quality/fidelity [18, 236, 237], which as discussed in Chapter 5 can have a significant impact on multimedia content experiences. The current study was designed to compare how different loudspeaker based audio playback systems can convey spatial information to a listener, whilst they are engaged in playing a video game. After playing a section of a video game, a group of participants were asked to subjectively rate the spatial sound quality of the experience based on a list of spatial audio attributes. The list was based on those audio attributes identified in Chapter 3 commonly used for audio quality assessment. Preference scores were then used to infer the degree to which the QoE of one gameplay experience might have improved in comparison to another. The subjective rating methods used relate back to the *analytic* and *utilitarian* quality measurements introduced in Chapter 5.

The rendering formats compared were mono, stereo and 7.1 surround-sound, which were chosen as they best represent the rendering options available to the majority of video game players at the time of writing (see Chapter 4). In all three of these listening

conditions, game audio was played back using all the loudspeakers of a 7.1 surround-sound arrangement according to the ITU specification given in Section 3.3. The justification for this decision is elaborated on in Section 6.3.2. A pilot test was also done to compare regularly configured stereo (i.e. played back using two loudspeakers in-front of the listener) and 7.1 surround-sound. Only loudspeaker based rendering methods were compared for two reasons. Firstly, it was of interest to determine how a physical loudspeaker array would perform, before considering VHT (see Section 3.4) options for headphone playback. Secondly, very few video games currently support multichannel headphone based playback.

This chapter begins with the method, design and procedure of the pilot study comparing regularly configured stereo and 7.1 surround-sound. This is followed by analysis of results and a discussion of the pilot study. After this, the pilot study is reflected on to inform the method, design and procedure of the main experiment comparing mono, stereo and 7.1 surround-sound played back over all the channels of a standard 7.1 loudspeaker arrangement. Results are then analysed and the overall process is discussed. The outcome of this experiment will be useful in determining how important multichannel game audio is. The test design is also novel in that there are few examples of similar studies in the literature using a real video game as experimental stimuli. Therefore the outcome will also help to inform future test designs by determining whether the video game used as stimuli is appropriate.

6.1 Research Question

The research question considered for the experiment presented in this chapter was as follows:

> *Will 7.1 surround-sound be perceived to have higher spatial sound quality than mono or stereo and will it be most preferred in the context of playing a video game?*

6.2 Pilot Study

It was important to conduct a pilot study in order to check the methods and materials used in the main experiment presented in Section 6.3. The pilot was also done to find if

the objective differences between stereo and 7.1 surround-sound are perceptible whilst engaged in a video game. If there was found not to be a difference, then this would require a redesign of the overall test and methods used, thus informing the main experiment. It was also useful to only compare two listening conditions, rather than the three considered in Section 6.3, as this allowed the entire process to be carried out more quickly.

6.2.1 Method

Participants played a video game with the audio played back to them either as a stereo or 7.1 surround-sound mix. All participants played the game using both experimental conditions. Upon completion of each game session, a set of spatial audio attributes were rated by each participant on a basic audio quality scale (see Section 6.2.3). Once both sessions had been completed, participants were asked to state which of the conditions was preferred and to what extent using a paired comparison scale.

6.2.2 Pilot Study Design

For the pilot test, two experimental conditions were compared: stereo and 7.1 surround-sound, both played back over loudspeakers and configured according to the diagrams given in Sections 3.2 and 3.3 respectively. These amounted to two independent variables for the study, the first being a stereo mix of the game audio played back over two active loudspeakers, and the second being a 7.1 surround-sound mix of the game audio played back over eight active loudspeakers (this included a subwoofer for the LFE channel). The two dependant variables were the perceived basic audio quality of a set of spatial audio attributes (given in Section 6.2.3), and the degree of preference for one of the two conditions.

Participants

A total of 6 participants took part in the pilot study, of which 4 were male and 2 were female. All who took part were aged between 18 and 30, and were recruited through the *Department of Electronic Engineering* at the *University of York* via email.

6.2.3 Pilot Study Procedure

All participants played through the video game stimuli (see the following section) under both experimental conditions which were not revealed until the entire experiment had been completed. The order of exposure was alternated between participants to ensure that both conditions were played first an equal number of times. Participants were briefed on the attributes used to rate the sound quality of a session before being exposed to any of the two conditions. Spatial quality was rated after each experimental condition and preference was only given once the player had been exposed to both conditions. The questionnaire given to each participant is covered in Section 6.2.3.

Materials

It was important for both the pilot and main experiment (see Section 6.3) to choose a game capable of outputting audio using up to 7.1 surround-sound. *The Last of Us: Remastered* [198], developed by Naughty Dog for the PlayStation 4 (PS4), was used for this reason. Critically, the game has received praise for its use of audio in the wider game audio community and was at one point the most awarded game in history [199]. Also, the importance of audio is stated early in the game's narrative, where the player is encouraged to listen for potential threats in order to gain a tactical advantage over enemy non-playable characters (NPCs). These cues are further emphasised when

FIGURE 6.1: A screenshot from the introductory sequence of *The Last of Us: Remastered* where the player is only required to move through the scene with limited interaction.

listening to the game's audio over a 7.1 surround-sound loudspeaker system, potentially influencing the way in which the game can be played. It was therefore of interest to determine whether the potential advantage of using 7.1 surround-sound in the context of this game would have an impact on preference scores when compared with mono and stereo.

To ensure that playing the game would not overly distract participants from the audio rating task, it was also important to consider those aspects of the content relating to gameplay/interaction. Work by Zielinski, Rumsey, Bech, De Bruyn, and Kassier [241] suggests that the visual aspects of a game world, and the attention required to successfully interact with it, can have a significant influence on an individual's ability to rate audio quality. The introductory sequence of *The Last of Us: Remastered* was chosen for ease of playability, in an attempt to not distract participants from the audio rating task. The player is required to follow a fairly simple and linear path with clear instructions from in-game events and sequences. The majority of the audio cues are scripted and will not trigger until the player encounters a particular section, ensuring similar auditory experiences between different players on multiple play-throughs. There are also a limited number of times the player can actually fail during the play-through, where, even if the player does not properly achieve an objective, they are

FIGURE 6.2: The listening room used throughout the experiment. Loudspeakers not part of the defined 7.1 playback system were non-active during the test.

able to continue with minimal consequence/loading time. The very beginning of the scene also acts as a short gameplay tutorial by gradually introducing players to the various control systems used throughout the game, such as movement using the PS4 gamepad analogue sticks and interaction using various buttons on the face of the gamepad. It is important to note that this is a commercially available game meaning there was no control over in-game audio rendering and in-game events. The pros and cons of this decision are discussed in Section 6.6 and Chapter 8.

For both the pilot and main experiment the game was played on a Sony Playstation 4 connected via HDMI to an Onkyo TXNR838 AV Receiver. Six Genelec 8040As, one Genelec 8040B (centre channel) and a Genelec 7060B Active Subwoofer were arranged according to the ITU specification [19] for 7.1 surround-sound listening and connected to the appropriate audio outputs of the receiver. This allowed for both stereo and 7.1 surround-sound listening conditions to be output over the same physical system. The overall level for each condition was controlled by the receiver and set to a comfortable level for the duration of the experiment. Game visuals were presented using an Optoma HD200X projector. An office chair was positioned in the centre of the listening array for participants to be seated whilst partaking in the experiment. The listening room, see Figure 6.2, was surrounded by a thick absorbing drape with foam acoustic paneling above the listener. The extra loudspeakers above, below and to the side of listener that do not conform to the 7.1 surround sound speaker configuration were not active.

Questionnaire

As discussed in Chapter 5, a *utilitarian* quality measurement is used to gain an overall impression of the stimuli's perceived quality. *Analytic* quality measurements break the assessment down into lower level attributes/characteristics, as a way to more specifically determine the aspects of the stimuli that may serve to have a more significant impact on the overall quality judgment. Both of these concepts were used to form the questionnaire used in the current study. The analytic portion of the questionnaire relates to spatial audio attributes that are believed influence judgements on audio quality. A preference score was used to gain an overall judgement between the experimental conditions. The spatial audio attribute list and questionnaire given to participants are presented in Appendices A.3 and A.6 respectively.

The list of spatial attributes is formed of some of the more commonly mentioned attributes from the literature [46, 116, 117, 119] with simplified definitions/descriptors to ensure understandability over a wide range of participants. These attributes are introduced in detail in Chapter 3. Participants were required to play a relatively long portion of a video game, therefore fewer attributes, with definitions, would be easier to remember, hence the the shorter definitions given below [1]. Before playing the game, participants were also able to confirm their understanding of the attributes with the principle investigator.

Localisation Accuracy (LA): Refers to how easy it is to identify the direction in which a sound source is originating. There should be good agreement between the visual location of an object/character in the game world and the sound it emits.

Distance Accuracy (DA): Refers to the perceived distance of sound sources. There should be good agreement between a sound source's perceived distance and the position of its related in-game object.

Sense of Depth (SoD): Refers to the perceived front-back definition of the sound scene and the sound sources within it. A scene with a good sense of depth will help to create a sense of auditory perspective.

Sense of Width (SoW): Refers to the perceived left-right definition of the sound scene and the sound sources within it.

Envelopment (Env): Refers to the extent to which the player feels surrounded by the sound presented in the scene.

Definition (Def): Refers to how multiple sound sources heard at the same time can be clearly identified and separated from one another.

The scale for basic audio quality (introduced in Section 5.1.3) was used by participants to rate each audio attribute. For preference, the paired comparison method detailed in Section 5.1.2 was used as it is an appropriate method when comparing only two experimental conditions at a time. Both the basic audio quality and paired comparison preference scales are given in Appendix A.6.

[1] These definitions have been revised for the purposes of this thesis, therefore some wording may vary to the questionnaires presented in the appendices.

6.2.4 Pilot Study Analysis of Results

The basic audio quality scores for the spatial audio attributes, and preference scores, were found to be non-normally distributed using a Kolmogorov-Smirnov goodness-of-fit test [242], therefore comparisons were made between pairs of conditions using sign-rank tests to check for any significant differences in the scores given. If the significance value (the p column in all preceding tables) output from the test is below the significance level of 0.05, this suggests that the difference in rating between the two conditions is significant and not by chance, and will be considered to be a rejection of the null hypothesis [242, 243]. Two null hypotheses were considered for the analysis based on the spatial attribute scores and preference ratings:

- There is no statistically significant difference between the spatial quality ratings given for each listening condition.
- There is no statistically significant difference between the preference ratings given for each listening condition.

For clarity, a rejection of the null hypothesis is represented by a value of 1 in the column h of the tables used to communicate the output of the statistical tests used throughout. When there is a significant difference, observing the relevant box-plot reveals whether the difference is due to one condition being rated overall higher or lower than the other. The same process for comparison was done for both the spatial attribute and preference analyses. The data and Matlab scripts for analysis can be found on the attached data CD, following the index in Appendix D.1.

The effect size is another statistical tool that is a standardised and objective measure used to determine how much the experiment explains the variance in results. If there is a significant difference, i.e. there is variance between the two conditions, the effect size (r) tells us how much the experiment actually effected that variance. An effect size of 0 means the experiment had no effect on the variance whereas 0.50 signifies a large effect, 0.30 a medium effect and 0.10 a small effect [242]. The effect size is defined as:

$$r = \frac{Z}{\sqrt{N}} \tag{6.1}$$

where r is the effect size, Z is a z-score produced from the sign-rank test, and N is the number of observations made (this includes multiple observations made by the same participants.). Ideally, to fully reject the null hypothesis it is necessary to observe a significant difference and a large effect.

Perceived Spatial Quality

An overall score for spatial quality was generated for each participant by summing the 6 individual attributes scores given for each condition. For each participant, this gave an overall spatial quality score between 0 and 30, for the two conditions they were exposed to. The closer the score is to this maximum value of 30, the more it suggests a higher overall perceived spatial quality for the respective condition.

The output of the sign-test analysis for the comparison of overall spatial quality is given in Table 6.1. [p = 0.041] shows that there was a significant difference between the two conditions and r is greater than 0.5, meaning there was a large effect. This implies that there is a good chance the difference in overall spatial quality was as a result of the loudspeaker configuration used. By observing the median lines on the boxplot in Figure 6.3, it can be seen that the overall spatial quality of 7.1 was, on average, higher than that of regular stereo.

Conditions		Median		T	p	z	r	h
Reg St.	7.1	20	23	0	0.041	-2.041	0.589	1

TABLE 6.1: Sign-test output for the comparison of overall spatial quality between regularly configured stereo and 7.1 surround-sound.

Preference

In regards to the preference scores, there was a statistically significant difference between the two conditions (Table 6.2). By inspecting the distribution of preference scores given in Figure 6.4, it can be seen that there is a clear difference between the two conditions, with the median line for surround being significantly higher than that of regular stereo and the large effect size [$r = 0.589$] suggests that this variance was due to the experimental conditions.

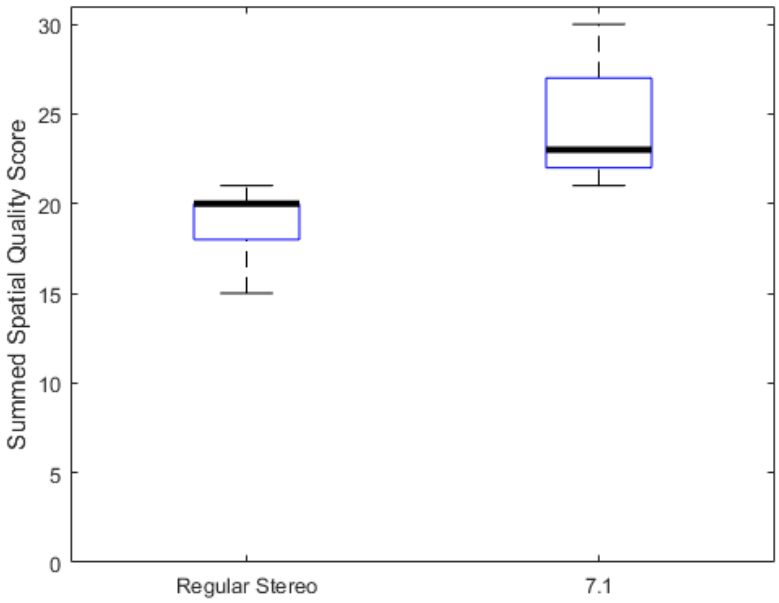


FIGURE 6.3: Boxplot showing the summed spatial quality scores for the control group conditions.

Conditions		Median		T	p	z	r	h
Reg St.	7.1	-3	3	0	0.041	-2.041	0.589	1

TABLE 6.2: Comparison between preference ratings for regular stereo and 7.1 surround-sound.

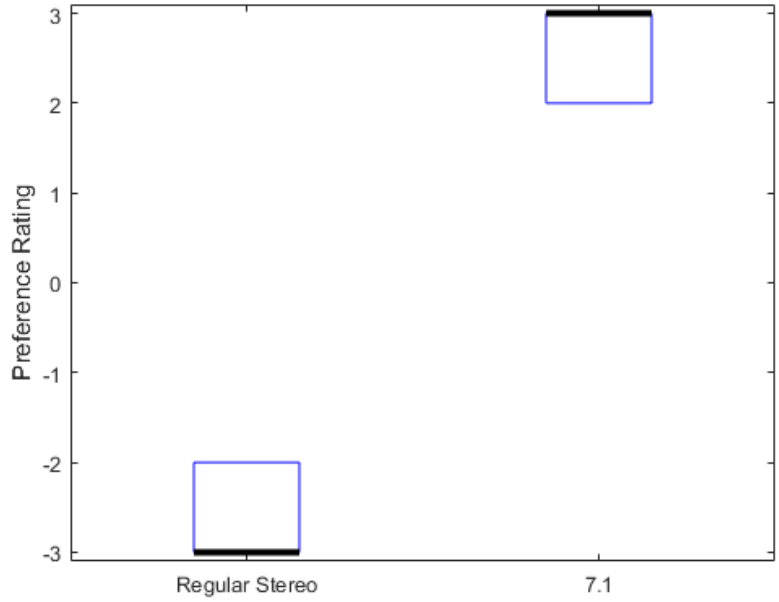


FIGURE 6.4: Boxplot showing the distribution of preference scores for the pilot group.

6.2.5 Pilot Study Discussion

The analysis for the summed spatial attribute scores provides good argument for the assumption that 7.1 surround-sound will be perceived to have higher spatial quality than regular stereo, output over two loudspeakers, even whilst playing a video game. The result was somewhat expected since the attribute definitions given to participants prior to each game session deliberately favoured a loudspeaker configuration able to render audio all around the listening space.

The preference results analysis gives the first indication that higher spatial quality is preferred when playing a video game, and is comparable to those results observed by Choisel and Wickelmaier [126], [215]. In addition, based on using preference as an indication of QoE, it can be inferred that the experience of the player was improved as a result of using surround-sound game audio over regular stereo.

There is however a possibility that the differences between the two conditions may have stemmed from the fact that the stereo condition made us of only two active loudspeakers in-front of the player, whilst the 7.1 condition had seven loudspeakers surrounding the player, with an additional sub-woofer. This may have therefore influenced the scores given by participants, due to the fact that there was an obvious difference in the number of active loudspeakers between conditions. Participants were aware that the experiment would involve the use of multichannel audio to some extent, potentially biasing their opinion in favour of 7.1 surround-sound. The listening conditions used in the main experiment were therefore modified so that regardless of the number of audio channels output from a particular listening format, all of the loudspeakers that make up a 7.1 surround-sound array would be active in an attempt to mitigate this potential bias. This process is discussed further in Section 6.3.2.

Overall, the pilot study was considered a success, in that the implemented materials, method and design yielded a clear difference between the two conditions, based on statistical analysis. For this reason, the main experiment followed the same methods and procedure using modified experimental conditions and a greater number of participants.

6.3 Main Experiment

The pilot test provided a good indication that the number of discrete audio channels used to present game audio had a positive influence on the perceived spatial quality of the game session, as well as preference. The main experiment expanded on the pilot by comparing three listening conditions instead of two. Other than the extra condition, the main difference in this experiment was that for the mono and stereo conditions, the audio channels were duplicated and output to all of the available loudspeakers in a 7.1 surround-sound arrangement. The rationale for this decision is discussed further in Section 6.3.2.

6.3.1 Method

A similar method was used to the pilot the study presented in Section 6.2. Those similarities are summarised below:

- Participants played a portion of a game twice under two different experimental conditions.
- After each game session, spatial audio attributes were rated on a basic audio quality scale.
- After the completion of both game sessions, the extent to which one condition was preferred over the other was expressed on a paired comparison scale.

The main difference in the method was that participants were not exposed to all of the experimental conditions, which was the case in the pilot. This was because a total of three listening conditions were considered for the main experiment. Instead, each participant was only exposed to two of the three conditions in order to reduce the time needed of each participant, as well as mitigate any potential learning bias after being exposed to the same material three times.

6.3.2 Experimental Design

Three playback conditions were used in the experiment: mono, stereo and 7.1 surround-sound. These three conditions were chosen as they are common game audio rendering methods used in commercial video game content (see Chapters 3 and 4). Although it

might be expected that other multichannel formats, such as higher-order Ambisonics, wave-field synthesis or higher channel count surround-sound systems might provide better spatial quality, these are formats that are not currently used in video game content hence their omission from this study.

As introduced in Section 6.2.5, there was a concern that results might succumb to some bias, due to it being easy to derive each listening condition based on the number of active loudspeakers. Therefore, in an attempt to keep the test blind, all 8 loudspeakers used in a 7.1 surround-sound arrangement were active for all three of the conditions. For the mono condition, a mix-down to mono of the game audio was output at an equal level from all of the loudspeakers. This is what is often referred to as 'full', or 'big', mono. For the stereo condition, audio intended for the left channel was output from the three loudspeakers of a 7.1 arrangement positioned to the left of the listener. The same was done with the right channel for the right-hand loudspeakers. The centre loudspeaker output a sum of the stereo channels. These conditions will be referred to as Big Mono (BMo) and Big Stereo (BSt) for the remainder of this chapter. For all the conditions, the routing of the individual audio channels to a designated loudspeaker can be found in Table 6.3. The loudspeaker angles used in the table relate to those illustrated Figure 6.5, representing a 7.1 surround-sound loudspeaker arrangement. The BMo and BSt rendering was handled by an external audiovisual (A/V) amplifier (see Section 6.2.3).

Loudspeaker Angles	Channel Allocation		
	7.1	BSt	BMo
0°	C	R + L	M
30°	R	R	M
90°	RS	R	M
135°	RBS	R	M
−135°	LBS	L	M
−90°	LS	L	M
−30°	L	L	M

TABLE 6.3: The allocation of channels to the angles of loudspeakers in the 7.1 surround-sound array. For *Big Stereo*, L and R correspond to the down-mixed left and right channels of the game audio, and M corresponds to the mono down-mix used for *Big Mono*.

Three independent variables were therefore considered for the main experiment. The first was a mono mix-down of the game audio played back at an equal level over seven loudspeakers. The second was a stereo down-mix split between seven loudspeakers. The third was the original 7.1 surround-sound mix used in the pilot study. The dependant

variables remained the same as those considered in the pilot study: spatial sound quality and preference.

Participants

21 participants took part in the main experiment (17 male and 4 female). 20 of these participants were aged 20-35 with one over 50. For this part of the test, participants were split evenly into the three groups outlined in Section 6.4, giving 7 participants per group and a total of 14 for each listening condition. All of these participants were different to the pilot study.

6.4 Experimental Procedure

Due to the length of the chosen scene from *The Last of Us: Remastered* (approximately 12 minutes) it was decided that participants would only be exposed to two of the three listening conditions, significantly reducing the amount of time required of each participant, whilst also reducing the risk of any learning effects that may occur after three play-throughs of the same content. For example, after already playing the game

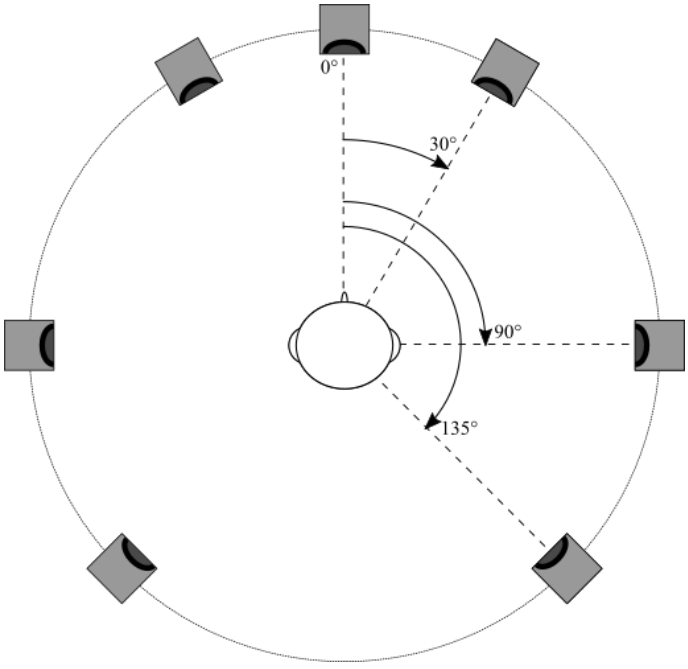


FIGURE 6.5: The loudspeaker angles used in all three listening conditions. The 7.1 arrangement is based on the standard given in ITU-R BS: 775 [86], with angles being symmetrical to the left and right of a front-facing listener. The channel to loudspeaker allocations for the BSt and BMo conditions are given in Table 6.3.

twice players, may become more familiar with in-game events, therefore skewing their reactions. Also, the paired comparison method used for preference ratings works best with only two experimental conditions. Participants were therefore assigned to three separate groups (A, B and C in Table 6.4), each of which were exposed to two of the possible three listening conditions:

Participant Group	Listening Conditions	
A	BM	BSt
B	BM	7.1
C	BSt	7.1

TABLE 6.4: Table showing the allocation of experimental conditions for the three participant groups, A, B and C.

Within each group, the order of exposure to the allocated conditions was alternated with each successive participant. Each participant received an experiment pack (see Appendix A) containing an information sheet, consent form, questionnaire, spatial attribute list with descriptors, a diagram of the game's control scheme and a summary of the in-game events to expect during the play-through (with approximate time markers). Before starting the experiment it was explained to participants that they would be required to concentrate on the specific spatial audio attributes outlined in the list, and these were further defined by the author to ensure a consistent understanding of the attributes between participants.

6.5 Results Analysis

Sign-tests were used again to compare pairs of listening conditions in order to keep the analysis consistent with that of the pilot study. It is however not possible to compare three listening conditions simultaneously using a sign-test, therefore analysis was done between every possible pair of the three conditions. Therefore, BMo was compared to BSt, BMo was compared to 7.1 and BSt was compared to 7.1. In the same way as the pilot, the overall spatial quality given by a participant was first compared between conditions. In addition to this the individual spatial attributes were also compared between conditions, without first summing then. Finally, preference was compared in the same way as the pilot.

6.5.1 Overall spatial sound quality

For the three comparisons, the output of the sign-tests are given in Table 6.5. From the analysis, it was possible to reject the null hypothesis for the comparisons between BMo and BSt, and BMo and 7.1 surround-sound. This shows that there was a statistically significant difference in the overall spatial quality scores in both cases (see Table 6.5). The median scores for both BSt and 7.1 surround-sound were both significantly higher than that for BMo [$Median = 14.5$], implying the perceptual spatial quality of game audio played back in this way was considered to be low.

Conditions		Median		T	p	z	r	h
BMo	BSt	14.5	24	2	0.016	-2.405	0.455	1
BMo	7.1	14.5	23.5	1	0.003	-2.94	0.556	1
BSt	7.1	24	23.5	8	0.79	0.267	0.051	0

TABLE 6.5: Comparisons by sign-tests between the overall (summed) spatial quality of the three listening conditions. There is a statistically significant difference for all comparisons other than that between BSt and 7.1 surround-sound [$p = 0.729$], suggesting the spatial quality of both was perceptually similar.

Unexpectedly, the null hypothesis could not be rejected for the comparison between BSt and 7.1 surround-sound, due to the significance level being [$p = 0.79$]. The result shows that there is no clear difference in the ratings given between the two, which is further evidenced by the similarity in the positions of the medians for BSt and 7.1 surround-sound (24 and 23.5 respectively), shown on the boxplot in Figure 6.6.

6.5.2 Individual attribute scores comparison

The overall spatial quality scores analysed so far were obtained by summing the scores of the six individually rated spatial attributes, *depth*, *distance*, *localisation*, *definition*, *envelopment* and *width*, providing a good high-level indication as to the performance of the playback system in this regard. However, it may have been that certain attributes were conveyed better in one condition than another, and others were conveyed similarly, therefore skewing the overall quality score. For this reason, the results presented in this section are based on analysing the individual attribute ratings, given between 1 (Poor) and 5 (Excellent). As per the previous analysis, this was done by comparing the attribute ratings by pairs of conditions with sign-tests, therefore giving three sets of results for every possible comparison.

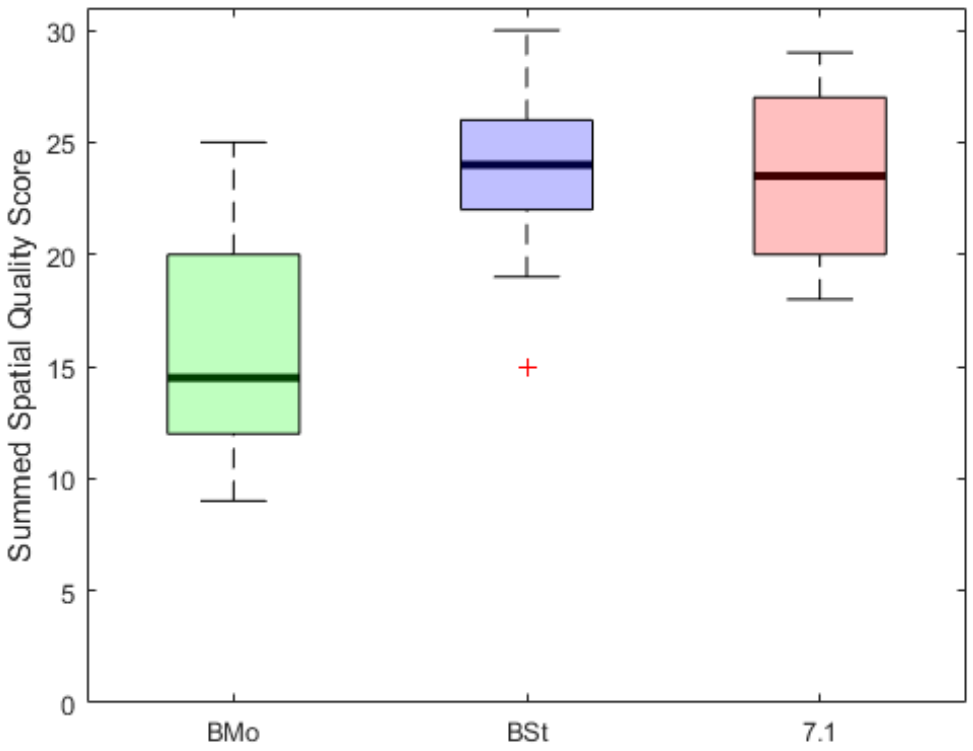


FIGURE 6.6: The distribution of summed spatial attribute scores for the three listening conditions BMo (green), BSt (blue) and 7.1 surround-sound (red). The summed scores are highest for BSt and 7.1 surround-sound, and lowest for BMo.

For the comparison between BMo and BSt, there was a statistically significant difference for all the attribute scores, as presented in Table 6.6. The median scores and boxplot in Figure 6.7 suggest that this difference was due to the attributes being rated consistently higher in the BSt condition, than BMo. Also, the large effect sizes (> 0.5) observed in the r column of Table 6.6 for *distance* and *definition* provides a strong indication that the difference between ratings was largely down to the listening condition used.

A similar comparison can be observed for some of the attribute scores between BMo and 7.1 surround-sound, as presented in Table 6.7, where the difference in ratings for the *distance*, *localisation* and *definition* attributes are significantly different. Again, this significance is a result of those BMo attributes being rated consistently lower than those for 7.1 surround-sound. The *depth*, *envelopment* and *width* attributes also have a higher median rating for 7.1 than BMo (3, 3.5 and 3.5 respectively), although the differences are not statistically significant.

As with the summed attribute comparison, there was no statistically significant difference between all of the BSt and 7.1 surround-sound attribute ratings. This is shown in Table 6.8 where for all the attribute ratings the significance values in the column marked p far exceed the significance level. The median scores are high relative to those for BMo, suggesting that both BSt and 7.1 surround-sound were perceived to have high spatial quality by the majority of participants, but were not significantly different to one another. Referring to the boxplot in Figure 6.7 there is a clear difference in the distribution of quality scores for both BSt and 7.1 (the Blue and Red plots respectively) in comparison to BMo, in that the Green plots representing BMo are overall much lower on the scale.

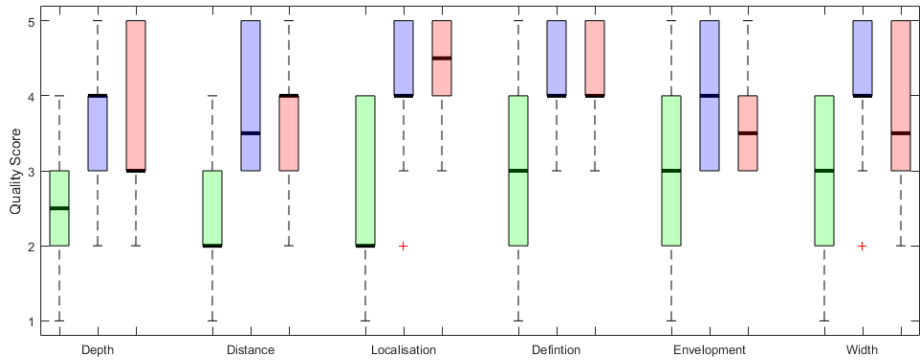


FIGURE 6.7: Boxplot showing the distribution of quality ratings for the individual spatial attributes given by participants to BMo (Green), BSt (Blue) and 7.1 surround-sound (Red). In general, the percieved quality of the individual attributes is significantly lower for BMo than for the other two conditions.

Attribute	Median (Mdn)		T	p (.05)	z	r	h
	BMo	BSt					
Depth	2.5	4	1	0.009	-2.598	0.491	1
Distance	2	3.5	1	0.006	-2.774	0.524	1
Localisation	2	4	1	0.009	-2.598	0.491	1
Definition	3	4	1	0.006	-2.774	0.524	1
Envelopment	3	4	1	0.027	-2.214	0.418	1
Width	3	4	1	0.009	-2.598	0.491	1

TABLE 6.6: Comparisons by sign-test for the individual spatial attribute scores between BMo and all BSt. The values of 1 in column h indicate that the quality of each attribute was significantly different between the two conditions, with the medians suggesting this was lowest when listening over BMo.

Attribute	Median (Mdn)		T	p (.05)	z	r	h
	BMo	7.1					
Depth	2.5	3	2	0.07	-1.809	0.342	0
Distance	2	4	1	0.006	-2.774	0.524	1
Localisation	2	4.5	0	0.001	-3.328	0.629	1
Definition	3	4	0	0.002	-3.175	0.6	1
Envelopment	3	3.5	2	0.07	-1.809	0.342	0
Width	3	3.5	2	0.114	-1.581	0.299	0

TABLE 6.7: Sign-test output for the comparisons of individual attributes scores between BMo and 7.1 surround-sound. The analysis suggests that there was a significant difference in the scores given for all of the spatial attributes other than *depth*.

Attribute	Median (Mdn)		T	p (.05)	z	r	h
	BSt	7.1					
Depth	4	3	5	1	0	0	0
Distance	3.5	4	5	1	0	0	0
Localisation	4	4.5	1	0.371	-0.894	0.169	0
Definition	4	4	2	1	0	0	0
Envelopment	4	3.5	6	0.752	0.316	0.06	0
Width	4	3.5	5	1	0	0	0

TABLE 6.8: Sign-test analysis for the attribute scores for BSt compared with 7.1 surround-sound. The 0s in the *h* column imply there was no statistically significant difference in any of the attribute ratings between the two conditions.

6.5.3 Preference

The output for the preference analysis is given in Table 6.9. A plot of the distribution of the ratings for the three conditions is given in Figure 6.8. The analysis shows that listening condition had a significant impact on the preference results for the comparisons between BMo/BSt and BMo/7.1 surround-sound, with the actual significance values being 0.003 and 0.002 respectively. In Figure 6.8 it can be seen that BMo received notably lower preference scores than the other two conditions, and both of the effect sizes (the first two values in column *r* of Table 6.9) are greater than 0.5, which is large effect.

The analysis also shows that there was no significant difference in the preference ratings between BSt and 7.1 surround sound (the bottom line of Table 6.9). Both were rated highly relative to BMo, as indicated by the respective median values, reflecting the analysis of the spatial quality scores. As with the spatial quality scores this implies that participants found it difficult to distinguish the two conditions, and/or participants applied the rating randomly.

Conditions		Median		T	p	z	r	h
BMo	BSt	-3	3	1	0.006	-2.774	0.524	1
BMo	7.1	-3	2	1	0.006	-2.774	0.524	1
BSt	7.1	3	2	5	1	0	0	0

TABLE 6.9: Results suggest a significant difference between all the comparisons other than the one between BSt and 7.1. This suggests neither was preferred more than the other, however on average both conditions received higher preference scores than BMo.

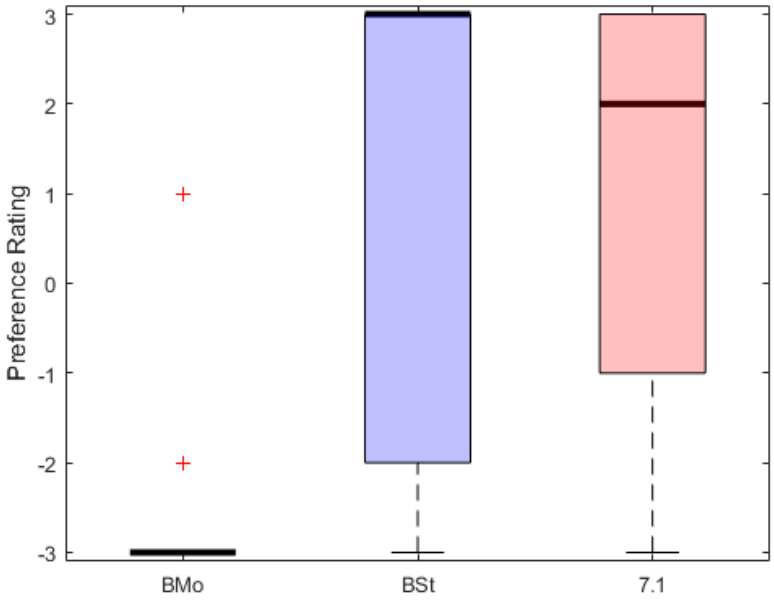


FIGURE 6.8: Boxplot showing the distribution of preference ratings for the BMo, BSt and 7.1 listening conditions.

6.6 Discussion

The results analysis shows that the overall spatial quality scores given by participants were higher for both 7.1 surround-sound and BSt in comparison to BMo. This was expected, since it is difficult to effectively convey the majority of the spatial attributes considered in this experiment over a single mono channel, as discussed in Chapter 3. However, analysis also showed that the difference in overall spatial audio quality between 7.1 surround-sound and BSt was not statistically significant. From this, it can be inferred that the overall spatial quality for the two conditions was perceptually similar, but significantly higher than BMo. This also implies that some participants potentially found it difficult to distinguish 7.1 from BSt, resulting in them allocating similar quality ratings for each attribute, or applying them randomly. For the presented study, this

might suggest that the number of active loudspeakers used for playback has more of an influence on perceived spatial quality than the number of discrete audio channels.

However, BMo was also output from all the loudspeakers of a 7.1 surround-sound system and was not perceived to have high spatial quality, potentially because no panning or separation of sound sources could occur over the one duplicated audio channel. The panning between the left and right channel will have been exaggerated to some extent in the BSt condition by the fact that the they were output from all the loudspeakers of a 7.1 arrangement. This extreme panning may have been perceived by listeners to be more spatial than regular stereo, resulting in an overall more positive spatial quality rating. This is reflected in a comment given by Participant 6 in Appendix A.8, who felt that sound sources were 'excessively spread out' in the BSt condition. It is important to note that BSt will naturally feel enveloping, since audio is physically output from all around the listener, even if the spatial information is not completely accurate in respect to the games visual feedback. One participant (Participant 6 in Appendix A.8) even felt that sounds were easier to localise in BSt than in 7.1 surround-sound. Visual stimuli can have significant effects on the perception of spatial attributes of audio stimuli, especially with respect to sound source localisation [41], [244]. It was stated by one individual (Participant 5 in Appendix A.8) that they could hear the sound of a helicopter overhead, even though loudspeakers mounted above the listener were not used in any of the three listening conditions. This extreme panning idea in BSt could be further investigated by testing the localisation accuracy of a participant using the system in comparison to a system with a higher number of discrete audio channels, like 7.1 surround-sound.

A large effect size was found for *localisation* and *definition* for both 7.1 surround-sound and BSt, which can be inferred as BMo being consistently outperformed for these attributes. Again, this is logical since it is difficult to separate out sound sources on only one audio channel, other than through amplitude manipulation, which in a game soundtrack may result in the definition between sources being lost. The same logic applies to localisation, where separated and discrete sound sources around the listener are not possible over one audio channel. These results suggest some advantages to using stereo and surround-sound in a game situation, where the potential in separating sound sources among multiple channels is clearly perceptible by listeners/players. However, at this stage, it is not possible to conclude that this will make the game easier

or more enjoyable to play. There is also general agreement between the comparisons made for the individual attributes and with the overall spatial quality analysis for the same pairs of conditions, suggesting that summing the attributes scores gives a good general representation of the spatial quality of a listening system. In this case, this shows that the overall spatial quality will be defined by how high in quality the individual attributes are perceived to be, and in general if the overall quality is shown to be high, then it is likely that this will be reflected in the individual attribute ratings.

Overall, the statistical analysis suggested that a listening condition perceived to have high spatial quality is also most preferred when engaged in the game. This was especially clear with regards to the BMo condition, which consistently received the lowest spatial quality ratings and was also least preferred, reflecting findings presented by Berg and Rumsey [115] and Dewhirst et. al [21]. However, both the BSt and 7.1 surround-sound conditions received similarly high spatial quality ratings and neither was preferred significantly more than the other. This is true for both the summed spatial quality scores and the comparisons made between the individual spatial attribute ratings. This result is surprising, as it was expected that, due to the high potential for spatialisation, 7.1 surround-sound would be perceived to have higher spatial quality than that of BSt. This assumption was made based on the results gathered from the pilot study, where the spatial quality of regularly configured stereo was perceived to be much lower than that of 7.1 surround-sound and was also not preferred by the majority of participants.

Reflecting on the attribute list given to participants, for those individuals not already familiar with the terminology, the descriptors used may have been difficult to understand. This brings into question whether the attribute rating system used was appropriate when using a video game as test stimuli, especially for the length of the play-through. The game session was as long as it was because it allowed participants to play an entire level of a game through to completion. This ensured every participant engaged in as close to the same content as possible, without having full control over the in-game events and soundtrack. However, this may have compromised abilities to effectively rate the content, because during each game session every attribute and descriptor had to be remembered. Even though efforts were made to increase understandability over a wide range of participants, it is not clear as to whether or not

this was effective. For example, the distribution of attribute ratings for BMo in Figure 6.7 are, for the most part, quite wide, suggesting some participants rated them highly in the BMo condition. This might be because it was genuinely felt that the spatial quality was high for BMo, although it might also imply that the attribute descriptions are being misinterpreted between participants. This is a regular problem with subjective tests of this nature, and it might be the case that a more objective measure would be more appropriate for tests involving video games and interaction from participants. This way data can then be gathered in real-time, without disturbing the player, and such approaches are the main focus of Chapter 8. A positive outcome for the subjective rating system was that the same conclusions were made by analysing the summed spatial quality scores and the individual spatial attribute ratings, which is encouraging. This shows that summing the attribute ratings for each participant offers a good general representation of a sound system's spatial quality, without the need to analyse each attribute individually.

Although it was unexpected that BSt was considered to have similarly high spatial quality to 7.1 surround-sound, it is important that both were equally preferred. From this it can at least be inferred that playing a game in what might be considered as being a listening environment with an enhanced spatial quality, at least in terms of listener perception, will be preferred, and potentially offer a more fulfilling quality of experience. This has positive implications for those gamers who cannot invest in a full surround-sound system, where BSt could be a viable alternative. Rather than using the eight loudspeakers needed for 7.1 surround-sound, perhaps players could benefit from more compact systems capable of outputting BSt, providing a heightened sense of spatialisation and envelopment over regular stereo.

6.7 Summary

This chapter presented a listening test designed to determine the perceived spatial quality of three different loudspeaker systems (mono, stereo and 7.1 surround-sound) whilst an individual was engaged in playing a video game. Preference ratings were also taken to infer which system offered the best quality of experience. Results suggest that a listening system with perceptually high spatial quality is preferable whilst playing a

game. A pilot study comparing 7.1 surround-sound with regular, two-channel stereo, showed that, overall, surround-sound was preferred. However, stereo played back over seven loudspeakers yielded similar spatial quality and preference ratings to the 7.1 surround-sound condition. This result was unexpected, as it suggests that listeners found it difficult to distinguish the two playback methods, even though 7.1 makes use of eight discrete audio channels, whilst big stereo only uses two separate channels that are duplicated and played back over seven loudspeakers. In all cases, mono was considered to have poor spatial quality, and was also not preferable. The next chapter focuses on virtualised headphone equivalents of the playback systems used here, to find if similar comparisons can be made in terms of the spatial quality and preferences.

Chapter 7

Headphone Based Audio Rendering and Player Preferences

This Chapter presents a perceptual listening test designed to be a continuation of that given in Chapter 6 which showed that a multichannel loudspeaker system with high spatial quality was also preferred whilst playing a video game. However, the physical loudspeaker systems used in Chapter 6 are not necessarily appropriate/available to the majority of video game players. For example, to fully experience 7.1 surround-sound, a listener needs specialist equipment such as eight loudspeakers (including a subwoofer) and an amplifier to drive the loudspeakers and decode the format used for transmission (see Section 3.3.1). This can be problematic due to both the cost and space needed for multiple loudspeakers. Also, the consistency of loudspeaker placement between different listening spaces (i.e. living rooms) is questionable [17].

A more practical approach is to virtualise the loudspeaker positions of a surround-sound system for listening over a pair of stereo headphones (introduced as a virtual home-theatre (VHT) system in Section 3.4). This retains the channel separation of a 7.1 surround-sound system, whilst also ensuring a consistent listening experience between different users. However, the previous studies comparing VHT systems to more commonly used stereo down-mixes presented in Section 5.2.2 do not provide positive arguments for the use of VHT systems, although participants did not directly interact with a video game in any of these cases.

On-line reviews and articles covering VHT systems for video games often contradict the studies given in Section 3.4 in that headphone-based surround-sound rendering systems

are often praised for offering an enhanced gaming experience [11–16]. The advantages of VHT systems are often mentioned in these publications, most commonly referring to enhanced feelings of immersion and localisation. In addition to this, articles that review headsets generally give high praise to those that make use of virtual surround-sound, often considering it as a key criterion. The gaming survey by Goodwin [17] also gives reason to believe that video game players consider surround-sound to be an important factor in a game experience.

For these reasons VHT systems for gaming should not be ignored in the context of this thesis as there is clearly a consensus in the wider (non-academic) gaming community that they are worthwhile. The primary aim of this chapter is therefore to investigate how the use of interactive audio stimuli might impact the perceived spatial sound quality of headphone based multichannel rendering methods, and whether the results are comparable with those found in the literature for non-interactive stimuli [100, 238, 239].

7.1 Research Question

The overall research question considered for the study presented in this chapter was:

> *Will participants rate the spatial sound quality of a VHT surround-sound rendering higher than a stereo down-mix, and will the VHT rendering be preferred?*

This question is used to derive both null hypotheses given in Section 7.5.

7.2 Method

Participants were asked to play a video game whilst listening to the audio over a pair of headphones. The game was played twice by each participant; once whilst listening to a VHT rendering (see Section 3.4) of the surround-sound game audio output, and once whilst listening to a stereo down-mix (see Section 3.3.2) of the same audio material. After each exposure to the game, each participant rated a set of spatial audio attributes on a five-point basic quality scale. Once each participant had been exposed to both the VHT rendering and stereo down-mix, they stated which of the two was preferred and to what extent.

7.3 Experimental Design

Two experimental conditions were compared in the current study. The first was a VHT rendering of the surround-sound audio output from *The Last of Us: Remastered*, with head-tracking. The second was a stereo down-mix of the surround-sound game audio, without head-tracking. Both of these conditions made up the two independent variables for the experiment. Because head-tracking was only implemented for the VHT condition, it was not a separate independent variable in this study. The head-tracking was one component that made up the whole VHT system, therefore participants were not asked to specifically consider the presence or absence of head-tracking between the two conditions. The dependant variables were the basic audio quality of a set of spatial audio attributes (given in Section 7.4.2) for each experimental condition, and the degree of preference for one of the two conditions.

The experiment is similar to that presented in Chapter 6, the main difference being that mono was omitted as an experimental condition because it consistently received the lowest scores in the previous test. The differences between stereo and surround-sound were more ambiguous, warranting further investigation. Additionally, this meant only two playback conditions were compared, streamlining the overall test procedure and analysis of results. By having only two conditions, it was not necessary to split participants into separate groups to assess different pairs of playback scenarios, as was the case previously, thereby increasing the total number of participant responses.

7.3.1 Participants

A total of 18 participants took part in the experiment (3 female, 14 male and 1 non-binary). 16 of these participants were aged $20-35$, with the remainder being over 35. All participants were recruited from the *Department of Electronic Engineering* at the *University of York* via email. 9 participants had played some part of *The Last of Us: Remastered* at some point before this experiment, as they were involved in the experiment presented in the previous chapter. The experiment pack given to participants, including information sheet, event time-line and questionnaire, is given in Appendix B.

7.4 Experimental Procedure

As only two listening conditions were considered, all participants were required to play through the introductory sequence of *The Last of Us: Remastered* twice under each listening scenario. The order of exposure was alternated between participants, meaning half of the participants played the VHT condition first and half played the stereo down-mix first. The rendering methods were not revealed until the experiment was over, and the participant had fully filled in the questionnaire. It was also necessary to recalibrate the headtracker used to account for head-rotations (see Section 7.4.1) between each playthrough due to potential drifting issues. Participants were encouraged to move their heads freely during both game sessions. Upon completion of each playthrough, participants were asked to rate the quality of each spatial attribute. Preference was only given after a participant had been exposed to both listening conditions.

7.4.1 Materials

In this section, the needs of the system used to realise the two experimental conditions are outlined. *The Last of Us: Remastered*, played on a PS4, was used again as experimental stimuli, for the option of 7.1 surround-sound playback, however the game does not natively have a VHT version of the surround mix. It was therefore necessary to generate a custom VHT headphone rendering, which was done in Max/MSP [245] using the *Spatialisateur* (*Spat*$\sim$) object library provided by IRCAM [108]. The down-mix to stereo was also done using Max/MSP. It is important to note that the low frequency effects (LFE) channel of the game audio output was not included in either headphone rendering, hence the term 7.0 surround-sound will be used for the remainder of this chapter.

Headphone based VHT systems are used as a means to present surround-sound content to listeners who otherwise could not experience such content over a physical loudspeaker array. As discussed in Chapter 3, the effect is achieved by binaurally processing the separate audio channels extracted from a piece of multichannel content with sets of HRTFs [46]. For a headphone based VHT system, the HRTF measurements used correspond to the position of the loudspeaker from which the channel is intended

to be output. The Max/MSP patch used in this experiment can be found on the attached data CD, following the index in Appendix D.2.

A dominant issue with VHT surround-sound content is the prevalence of front-back confusions, occurring when sounds located directly in front or behind the listener provide similar interaural time and level differences [38, 41]. The resultant effect of this phenomenon is that sounds intended to be coming from in-front of the listener will appear to come from behind, and vice-versa. In a loudspeaker listening environment, a listener is able to correct these reversals by performing small left/right head movements. However, this is not possible when listening to a virtualised version of 7.0 surround-sound content over a pair of headphones, as the virtual soundstage will follow listener head movements. This is not ideal when visuals are presented using stationary monitoring equipment, such as a television set. As discussed in Chapter 3, one solution is to perform a compensatory rotation of the virtual soundstage in the opposite direction to the listener's head movement based on data obtained through head-tracking. It is well established that transformations, such as rotations about a vertical axis, can be applied to listening material encoded into Ambisonic B-format [110]. This was therefore done to the individual channels of the 7.0 game audio output before HRTF processing was applied.

VHT Rendering Process

Figure 7.1 is a block diagram showing the signal flow for the multichannel game audio from the PS4 output to *Spat*$\sim$ (running in Max/MSP) and then to headphones, for the VHT listening condition. The seven individual channels that make up the 7.0 surround-sound signal were first extracted from the HDMI port of the PS4 via an Onkyo TX-NR838 AV receiver. The outputs of the receiver (not including the LFE channel) were then patched through to Max/MSP, running on a Macbook Pro, via an RME Fireface UCX audio interface, for further processing.

After extracting the game audio, the first stage of processing converted the seven individual surround-sound channels to 3rd order B-format using a *Spat*$\sim$ Ambisonic panner object. Each respective channel was panned according to the loudspeaker angles for 7.1 surround-sound in ITU-R BS.775-3 [86], relative to a listener facing the centre channel (see Figure 3.3b). The game audio was encoded to 3rd order because as the

Ambisonic order increases, so does the spatial resolution of the encoded audio material, in that accuracy in localisation between sound sources is improved [112–114]. In using 3rd order, there is an increased likelihood that the individual channels will be perceived to originate from the correct locations around the virtual listening space, once the material was decoded and HRTF processing was applied.

Compensatory rotations of the encoded material were performed based on the direction in which the listener was facing. The horizontal angle in degrees, or yaw, of the listener's head was obtained using an EDTracker Pro [246], an inexpensive head-tracking unit designed to be used by video game players, mounted to the headphones used for playback. This angle was used in a *spat.hoatransform~* object, setup to rotate the encoded surround-sound channels about a horizontal axis. The object matrix multiplied all 16 channels of the 3rd order B-format signal based on a similar rotation matrix to the one outlined in Section 3.5, using the obtained head angle in degrees [1]. The object therefore output a version of the 16 channel B-format with the transformation applied. This rotation, performed in real-time, gives a listener the

[1]The exact rotation matrix is not defined in the *Spat~* documentation, however those presented in [106] for higher-order Ambisonics systems provide a good indication.

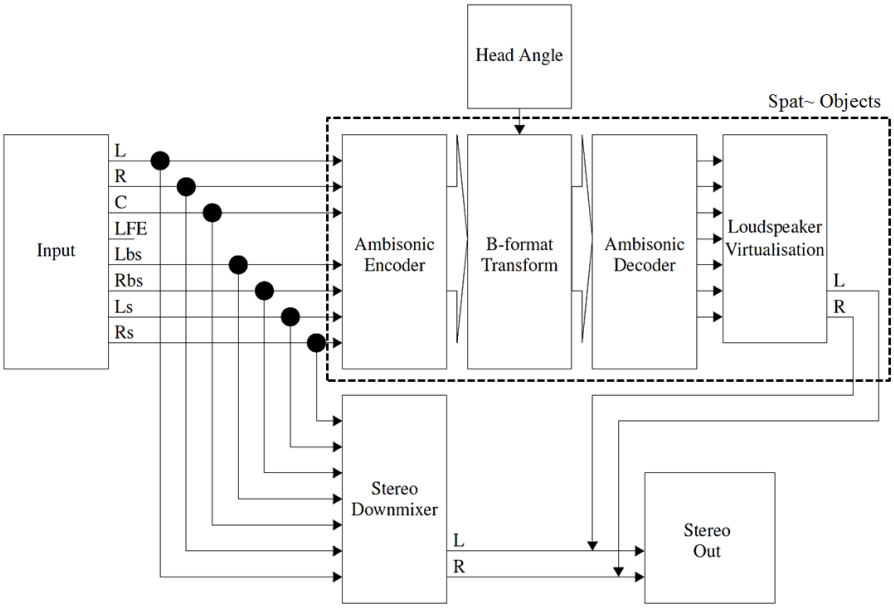


FIGURE 7.1: Block diagram showing the Max/MSP patch signal flow. The patch generates both a stereo down-mix and virtualised version of 7.0 surround-sound material to be presented to a listener over headphones. *Spat~* objects were used for Ambisonic processing and loudspeaker virtualisation.

impression that the virtual loudspeakers are stationary and do not follow head movements, as would be the case in a physical listening environment.

Once the rotation was applied, the 3rd order B-format was decoded back to the seven channels of the original surround-sound game output. For the system presented here, this was done by setting the decoder in *Spat*$\sim$ to a custom loudspeaker arrangement made up of seven angles that conform to the ITU standard for 7.0 surround-sound listening; 0°, 30°, 90°, 135°, -135°, -90°, -30° [2]. These were the same angles used for the physical surround-sound loudspeaker configuration in the experiment presented in Chapter 6 (see Figure 6.5). The individual channels were then processed using HRTF sets corresponding to the same loudspeaker angles in order to correctly simulate the loudspeaker positions for playback over headphones. This was done using a *spat.virtualspeakers*$\sim$ object set for 7.0 surround-sound. The object accepts the seven channels of a surround-sound signal, applying the HRTF sets, and outputs a left/right stereo signal that is played back over a pair of headphones. The default HRTF sets provided in *Spat*$\sim$ were used. There is research to suggest that binaural room impulse responses (BRIRs) are more appropriate for VHT rendering as they also contain the acoustic properties of the listening space, which helps to externalise the virtual loudspeakers thereby giving the impression of a more natural listening environment [46]. However, it is unclear as to whether the virtualising headsets already available to video game players take this into consideration. Therefore it was decided BRIRs would not be used for the virtualisation process in this experiment.

Stereo Mix-down Process

For the stereo down-mix playback condition, the same process outlined in Section 3.3.2 was used, not including the phase shift used in some Dolby systems. The phase shift is not given in the ITU specification for down-mixes to stereo [86], hence its exclusion from this experiment. The mix-down was generated in parallel with the VHT rendering using the same methods outlined previously in this section to extract the game audio surround output. This ensured the same audio material was processed for both experimental conditions. For both conditions, participants listened to the audio using a pair of Beyerdynamic DT 990 Pro stereo headphones. The experimenter was also able to

[2] Loudspeaker angles are presented in a clockwise orientation originating from a front/centre position with 0° elevation.

switch between the VHT rendering and the stereo down-mix using a button on the graphical user interface of the Max/MSP patch.

7.4.2 Questionnaire

The questionnaire used in the current study was similar to that used in Chapter 6. Therefore each playback condition was rated based on the spatial quality, determined by individually rated spatial attributes, and an overall preference score given after exposure to both conditions. The attributes considered were *localisation accuracy* (LA), *distance accuracy* (DA), *sense of depth* (SoD), *sense of width* (SoW) and *Envelopment* (Env). See Section 6.2.3 for full definitions of the attributes. *Sound source definition* was not been included due to its similarity with the description used for *localisation accuracy*. Attribute quality was rated on a 5-point numerical scale structured as: (1) Bad, (2) Poor, (3) Fair, (4) Good and (5) Excellent [218].

Preference was rated in exactly the same manner as in Chapter 6, using a 7-point paired comparison scale structured as: Strong preference for A, preference for A, slight preference for A, no preference, slight preference for B, preference for B, strong preference for B. The full questionnaire, with participant information sheet/consent form are given in Appendices A.6 and B.1 respectively.

7.5 Analysis of Results

The same analysis methods were used here as was done on the results gathered in Chapter 6, using sign tests and calculated effect sizes to determine any significant differences between the listening conditions. In Chapter 6, it was determined that comparisons between individual spatial attribute scores gave a good indication as to the overall spatial quality of a listening system. For this reason only the individual attribute ratings were analysed as it can be assumed that if the majority of attribute scores are high, then the overall spatial quality of that listening condition will also be high. All statistical analysis was conducted in MATLAB using the *Statistics and Machine Learning Toolbox*. The data and Matlab scripts for analysis can be found on the attached data CD, following the index in Appendix D.3.

	Median (Mdn)		T	p	z	r	h
	Down-mix	VHT 7.0					
Preference	-1	1	12	0.239	1.179	0.196	0
Localisation Accuracy	4	4	7	1	0	0	0
Distance Accuracy	3	4	7	0.343	0.949	0.158	0
Sense of Depth	3	4	9	0.267	1.109	0.185	0
Sense of Width	4	4	7	1	0	0	0
Envelopment	3	4	11	0.061	1.871	0.312	0

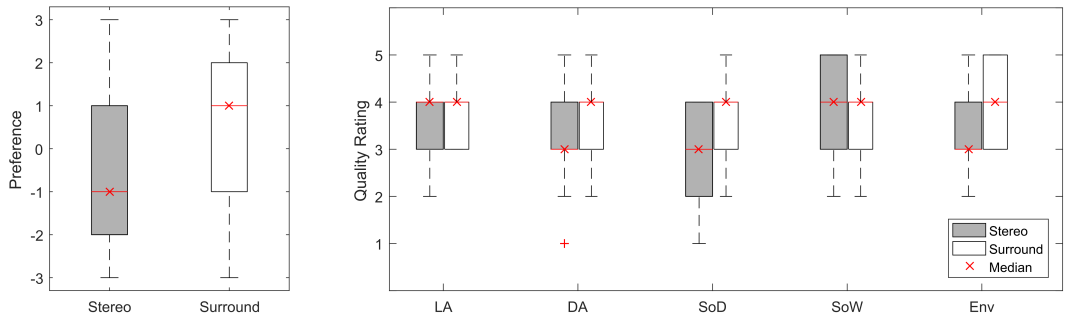
TABLE 7.1: Sign-test output for preference and spatial attribute quality ratings. T is the signed-rank and p is the significance value. The z score is used to determine the significance value (p) and the effect size (r). A value of 1 in the h column signifies a rejection of the null hypothesis.

Spatial attribute and preference ratings for each listening condition were first checked for normal distribution using a Kolmogorov-Smirnov goodness-of-fit test [242]. Participant responses were found to be non-normally distributed (non-parametric), therefore sign tests were used to check for significant differences in the data. The null hypotheses considered for analysis were:

- There is no statistically significant difference in perceived spatial quality between the two listening conditions.
- There is no statistically significant difference in preference between the two listening conditions.

Table 7.1 presents the output of the sign test analysis for the preference and spatial attribute ratings. A value of 1 in the column labeled h signifies a rejection of the respective null hypothesis at the chosen significance level of $p < 0.05$. The column marked r is the effect size, calculated according to [242] using the values from column z.

It was not possible to reject the null hypothesis for the preference ratings, as presented in Table 7.1. The significance value ($p = 0.239$) shows there was not a statistically significant difference between the preference scores given to each condition, implying neither listening condition was preferred significantly more than the other. The boxplot in Figure 7.2a shows the distribution of preference ratings for the two listening conditions, where it can be seen that although the median lines of the conditions are different to one another, the distribution of values is too wide for the difference not to be by chance.



(a) (b)

FIGURE 7.2: Boxplots showing the distribution of preference ratings (a) and spatial attribute quality scores (b) for the two listening conditions. Median scores are indicated with a cross for clarity.

The null hypothesis could also not be rejected for the spatial attribute quality scores given for *localisation accuracy*, *distance accuracy*, *sense of depth*, *sense of width* and *envelopment*, showing no statistically significant difference between the two listening conditions for any of the attributes considered. The boxplot in Figure 7.2b shows the distribution of quality scores given for the spatial attributes between the two listening conditions, where it can be seen that the median lines for almost all of the attributes are relatively similar, suggesting there was little to no perceptual difference between the conditions.

7.6 Discussion

The analysis of results showed that the virtual surround-sound (VHT) condition was not preferred over the stereo down-mix, and there was no significant improvement in the perceived spatial quality. The *sense of width* (SoW) and *localisation accuracy* (LA) attributes are particularly noteworthy, as a significance value of $p = 1$ implies that regardless of the rendering technique, there was no change in how each listening system was able to present these sensations to a listener. However, even though there wasn't any perceptual difference between the VHT rendering and the stereo down-mix, both received relatively high spatial quality scores, again suggesting that high spatial quality is preferable. This was observed in Chapter 6, where physical 7.1 surround-sound and big stereo were both perceived to have higher spatial quality than mono, but were not different from one another. For headphone rendering, the outcome was not entirely unexpected, as similar work presented by Pike and Melchior [238], Zacharov and Lorho [100], and Sousa [239] yielded similar results. The fact that a video game was used in

place of more traditional non-interactive material, does not seem to have had any significant impact on the perceived performance of the 7.1 system. It is therefore difficult to conclude that, when using a VHT system, a player's game experience can actually be improved when compared to a down-mixed stereo rendering.

Reflecting on the implementation of head-tracking, the process of performing rotations on a 3rd-order B-format version of the 7.0 surround-sound stream may have actually been detrimental to the overall quality of the render. Since the game audio had to be encoded and decoded using Ambisonic processing, the spatial resolution of the render will have been compromised in comparison to using all seven of the raw surround channels. It also isn't clear if the head-tracking was actually necessary for the VHT condition, although this was not formally tested. Since visuals were played on a stationary television monitor, players didn't necessarily need to move their head at any point during a play session. Even if the compensatory rotations were beneficial to the VHT rendering, with minimal head movement under the gameplay conditions it is possible that some participants may not have even noticed head-tracking had been implemented in the system. Head-tracking is far more appropriate for virtual reality (VR), where head mounted displays (HMD) are used for visual feedback, and players can use head movements to directly interact with a game, such as by influencing the camera view. Head-tracked audio systems, like the one presented in this chapter, are therefore becoming increasingly popular in VR gaming applications, rather than in more traditional gaming situations where the player is usually stationary, and looking straight ahead a television set or PC monitor.

For VHT systems, ideally unique head-related transfer function data-sets should be used for each participant, with some simulation of room acoustics to improve externalisation [46]. This may have resulted in the lack of perceptual difference between the conditions. The use of non-individualised HRTFs can significantly reduce an individual's ability to perceive spatialised aspects of audio material when listening over headphones [56]. In extreme situations this can result in the incorrect spatialisation of audio and undesirable timbral colourations. Unfortunately the use of generic/non-individualised HRTFs is representative of virtual surround-sound gaming headsets, due to the difficulties in collecting individualised measurements.

Even though attempts were made to improve the understandability of spatial attributes

between participants, there is still a possibility that participants did not fully understand the attributes they were asked to rate, thus impacting their ability to effectively rate the stimuli. For both this chapter and Chapter 6, it may have been appropriate to provide participants with auditory examples of each attribute under consideration to go along with the written description. Such examples are provided by Berg and Rumsey [247] and suggested in the Spatial Quality Inventory [116], although due to the dynamic aspect of video games, it would be difficult to create a fully representative set of audio example attributes based on what would be expected during the game session. It would perhaps be better to eliminate the subjective aspect of the test entirely, which is the main point of discussion in Chapter 8.

It is also difficult to say whether the multi-modal task of playing a video game served to have a positive or negative influence on the way in which participants were able to rate the listening conditions. As work by Zeilinksi suggests, participants may not be able to consistently rate audio content with interactive and visual elements [241]. With respect to the two experimental chapters already presented, participants had to play a relatively long section of the game (12-16 minutes) before they had a chance to give their opinions on the audio presentation. It is therefore not unreasonable to surmise that results, in both cases, may have succumbed to some bias, based on a listener's inability to only focus on the auditory aspects of the game session, or only being able to recall the most recent events that might not have been representative of the whole extract played. One solution would be to provide players with a training session, before the real test, using a listening mode not being assessed (or with the game audio turned off) to allow players to become used to the game controls and systems. This would reduce the risk of players learning the game, whilst also being asked to rate specific auditory attributes. Based on this, and the understandability of the spatial attributes used, it may be that listening tests based on subjective measures are not appropriate in experimental situations where participants are so engaged in the task, and instead, objective quantification might prove to be more reliable. This could include quantifying a player's score, or the time it takes to complete an in-game task. Data can then be collected in real-time, potentially reducing the cognitive load on participants. A novel method, taking the points of objectivity and player training into consideration, is presented in Chapter 8.

7.7 Summary

This chapter presented a subjective listening test designed to compare the perceptual differences between a VHT rendering and stereo down-mix of 7.0 surround-sound game audio, presented over headphones. This was based on spatial audio attribute ratings and preference scores. A custom VHT game audio rendering system was implemented in Max/MSP, using the *Spat*$\sim$ object library to implement head-tracking so that compensatory rotations could be applied for the virtual loudspeaker sound sources.

Analysis of results suggested that there was no statistically significant difference between VHT and stereo down-mixed game audio for this study. Also, neither system was preferred over the other, although both were perceived to have relatively high spatial quality in comparison to the low spatial quality observed for the mono condition in Chapter 6. The results have negative implications for the use of VHT systems in gaming, even though the general consensus among the gaming community is that surround-sound virtualisation over headphones is beneficial. It is also thought that the results in both this chapter and Chapter 6 may have been negatively influenced by the subjective measures used (i.e. the spatial attribute rating systems). The next chapter will attempt to eliminate this by introducing a novel objective experimental method. This will allow for data to be gathered in real-time from participants meaning that there is a reduced chance of engagement being broken during the game session, and results will not rely on them have to remember a list of attributes with descriptors.

Chapter 8

The Impact of Multichannel Audio on Player Performance

This chapter explores the impact multichannel rendering has on the performance of someone playing a gamified audio localisation task, using both objective and subjective metrics. If a player is able to better determine the location of an in-game event that might help them progress through the game, will this inform their gameplay decisions thus providing them with an advantage? This is considered for both loudspeaker and headphone based versions of stereo, 7.0 surround-sound and an octagonal loudspeaker array. Previous chapters have explored the use of subjective methods, in the form of sound attribute and preference rating scores, however, the variability in results (i.e. the standard deviation between different people) raised the question as to whether participants had similar understandings of the attributes being rated. Also, it is unclear as to whether the task of playing a game served to negatively impact the subjective scores given, in that participants may not have been able to fully focus on the audio rating task given to them. In measuring the performance of the player during gameplay, it eliminates the need for them to remember a list of attributes, or to break their engagement in order to complete a questionnaire. From this concept, a novel idea for a comparative listening test is derived, designed according to three core principles:

1. The player's objective is to locate as many sound sources as possible in a given time limit.

2. The player does not receive any visual feedback regarding the position of the sound source.

3. The player receives a final score determined by how many sound sources they found, and their in-game path to the sound source is recorded, giving two objective measurements relating to player performance.

The number of correct localisations gives a direct measure of how well a player performs in the game, which can be easily compared between the different experimental conditions. It is expected that scores will be lowest in the stereo condition because it will be more difficult to localise a sound-source when it is positioned to the side or rear of the listener. This is due to the lack of loudspeakers (real or virtual) at those positions. However, there is uncertainty based on the results given in Chapter 7, where the subjective qualities of virtual surround-sound and stereo rendered for headphones was similar. It is therefore of interest to assess whether the objective performance of a player in a task where more channels are advantageous will reflect those results from Chapters 6 and 7.

Two-channel stereo gives strong frontal phantom imaging due to the placement of the left and right loudspeaker at $\pm 30^{\circ}$ relative to the central listening position (see Chapter 3). However, imaging to the sides and rear is not possible due to the lack of loudspeakers at these positions. It is therefore expected that a listener would find it difficult to locate a sound anywhere but within the $\pm 30^{\circ}$ of the stereo pair. 7.0 surround-sound retains the left/right stereo pair but expands on it through the addition of 4 loudspeakers to the sides and rear of the listener. Also, the centre channel further defines imaging in the front quadrant. For this study the sub-woofer of 7.1 surround-sound ('.1' channel) was not included, as it is intended for further defining low frequency effects, hence referring to the surround-sound condition as 7.0.

As discussed in Chapter 3, the angles between the rear and side channels used in the 7.0 condition exceed the recommended 60° for stable imaging. For this reason, imaging is generally inconsistent at any position other than those represented by a physical sound source, i.e. a loudspeaker [89], [88], [90]. Therefore, even though player performance may improve over stereo as a result of the additional channels, the 7.0 arrangement is still not ideal for stable imaging at every point around the listener. Theile and Plenge propose an equally spaced arrangement of six loudspeakers to get a suitable 'all-around' effect, especially for sources intended to be perceived from $\pm 90^{\circ}$ [90]. This configuration was extended by Martin et. al [91] to an equally spaced octagonal array with a front

center speaker placed at 0° relative to the listener. The array was found to give relatively stable imaging around the listening space for amplitude-based panning algorithms. The conclusions drawn from these studies provide evidence that a listener's ability to successfully localise a sound will be influenced by the phantom image stability of the loudspeaker array used. However, none of these studies asked participants to directly interact with audio stimuli by playing a game. Therefore it is of interest to investigate whether similar comparisons can be made between different loudspeaker arrays, with varying degrees of phantom sound stability, in the context of an interactive game task. However, for the octagonal arrangement, the loudspeakers in front of the listener need to be spaced at a wider angle than those in the 7.0 and stereo arrangements if equidistant placement is to be achieved, with two loudspeakers placed at $\pm 90°$ for reliable lateral imaging. Therefore the trade-off in ease of localisation between more consistent imaging all around the listener, and the potential for higher resolution frontal imaging in 7.0 and stereo, is also of interest.

8.1 Research Question

The research question considered for the experiment presented in this chapter was as follows:

> *Will a player be more successful at a gamified localisation task if they are using a listening system with a higher number of discrete audio channels, and will the higher number of channels be preferred?*

8.2 Method

Participants were required to find the location of a sound source in a custom game environment, without being able to see a visual representation of it. This meant participants were only able to listen for the sound source. The game was played three times by each participant, each time using a different audio rendering solution. Half of the participants used loudspeaker-based solutions, whilst the other half used headphone-based solutions. Each playing session was measured according to how many sound sources a participant was able to find in a two and a half minute time period. This gave three scores for each participant, one for every listening condition

they had been exposed to. A subjective preference rating was also used to supplement the player scores, and as in previous chapters was used to infer the overall experience of the player after a game session.

8.3 Experimental Design

Two groups of participants took part in this study, Group A and Group B. The members of Group A had the game audio played back to them over loudspeakers, whilst the members of Group B listened over headphones. Within each group three listening conditions were compared, making three independent variables for each group of participants: a stereo down-mix, 7.0 surround-sound and an octagonal array of loudspeakers. For Group B, the 7.0 surround-sound and octagonal conditions were VHT renderings for headphone playback (see Section 3.4). Repeated-measures test designs such as this are susceptible to learning effects, in that participant results may be influenced through being exposed to the same program material multiple times. To reduce this risk, the order of listening conditions was counterbalanced within each group as suggested in [242]. With three listening conditions, this gives six sub-groups (1 to 6) within Group A (the loudspeaker group) and six sub-groups within Group B (the headphone group), see Table 8.1. For each group there were 24 participants (see Section 8.3.1), so for the 3 listening conditions this gave 4 participants in each counterbalanced sub-group. Furthermore, a training session was provided, as described in Section 8.4.4.

Sub-group	Condition		
1	Stereo	7.0	Octagon
2	Stereo	Octagon	7.0
3	7.0	Stereo	Octagon
4	7.0	Octagon	Stereo
5	Octagon	Stereo	7.0
6	Octagon	7.0	Stereo

TABLE 8.1: Counterbalanced sub-groupings for the three listening conditions within group A (loudspeaker playback) and group B (headphone playback).

Each game session lasted 2 minutes 30 seconds, with a 'Game Over' message and the player's score (i.e how many times the sound source was correctly located) being displayed on-screen at the end of each session. The number of correct localisations was output to a separate text file after each game session, giving each participant a final

score for each of the three listening conditions. The in-game path taken by each participant for each listening condition was also recorded in order to compare it to the shortest possible path. Once a participant had been exposed to all of the listening conditions within their group, they were asked to state which of the three conditions they preferred, and provide any comments regarding the experiment. Therefore, for each listening condition three dependant variables were measured: the player score, the in-game path taken by the player and the player's preferred listening condition. Participants were not made aware of any of the conditions prior to, or during, the test.

8.3.1 Participants

The experiment consisted of two groups (A and B) each comprising of 24 participants. Group A were exposed to the three conditions (stereo. 7.0 surround-sound and octagon) played back over loudspeakers, whilst group B were exposed to the headphone based equivalents of the same three conditions. For group A, 16 participants were male, 6 female, and 2 non-binary. For group B, 17 were male, 5 were female and 2 non-binary. Across groups A and B, all participants were aged between 18 and 40. All participants were recruited via email and had some affiliation with *The University of York*. Before participating, all potential participants were asked if they were familiar with using a gamepad to control a game. If not, they were asked not to participate in order to reduce the amount of time needed to learn the game's control system. All provided a signature to confirm their consent. Some participants were exposed to both the loudspeaker and headphone conditions, although they were rated on different occasions and more than 6 months apart. The experiment pack given to participants from both groups is given in Appendix C.

8.4 Experimental Procedure

This section outlines the localisation task participants were asked to complete and how it was implemented using a game-like virtual environment. The methods used to render stereo, 7.0 surround-sound and an octagonal array over both physical loudspeakers and headphones are then covered. It was decided early in the design process that a custom-made game environment would be used. The program material used in previous chapters was taken from a commercially available video game for

current-generation gaming consoles. However, it is not possible to access the source code of such content, making it difficult to determine the exact audio rendering methods used, beyond the way in which loudspeakers should be placed. The repeatability between participants is also questionable, along with potential learning effects that may occur due to multiple play-throughs of the same piece of game content. Creating a custom video game gave more control over the underlying mechanics/systems and the effectiveness of an octagonal loudspeaker array could be more easily explored.

8.4.1 Materials

The virtual environment and underlying systems for the localisation task were designed and implemented using the Unity game engine [248]. Sound spatialisation and rendering for the loudspeaker conditions were done separately in Max/MSP [245]. A single sound source was used in the game, the position of which changed as soon as it was successfully located by the player. The sound source was represented by a spherical Unity game object with a radius of 0.5 metres and its visual renderer turned off, ensuring that the source would be invisible to participants. The position of the sound source was always determined randomly within the boundaries of the game world, represented by a 20x20 metre square room. Random positioning was implemented so that players would not learn sound source positions after playing the game multiple times. The virtual room was comprised of four grey coloured walls and a floor and a ceiling to serve as a visual reference regarding the player's position within the game world, see Figures 8.1 and 8.2. According to Zielinkski et al. [241], visuals can distract significantly from an audio-based task, therefore visuals were deliberately simplified.

Players were able to navigate the game world through the eyes of a virtual avatar, using a control system similar to those found in the majority of first-person point of view games. The position and rotation of the avatar, within the boundaries of the game world, could be controlled by the player using the left and right joysticks of a standard Playstation 4 gamepad. This allowed for full 360° movement in all directions on a horizontal plane. The gamepad's 'x' button was used to trigger a simple `if` statement within the game's code to determine whether the player had successfully found the sound source. If, upon pressing the 'x' button, the player avatar was within the radius of the sphere representing the sound source's current location, the sphere would move to a random new location at

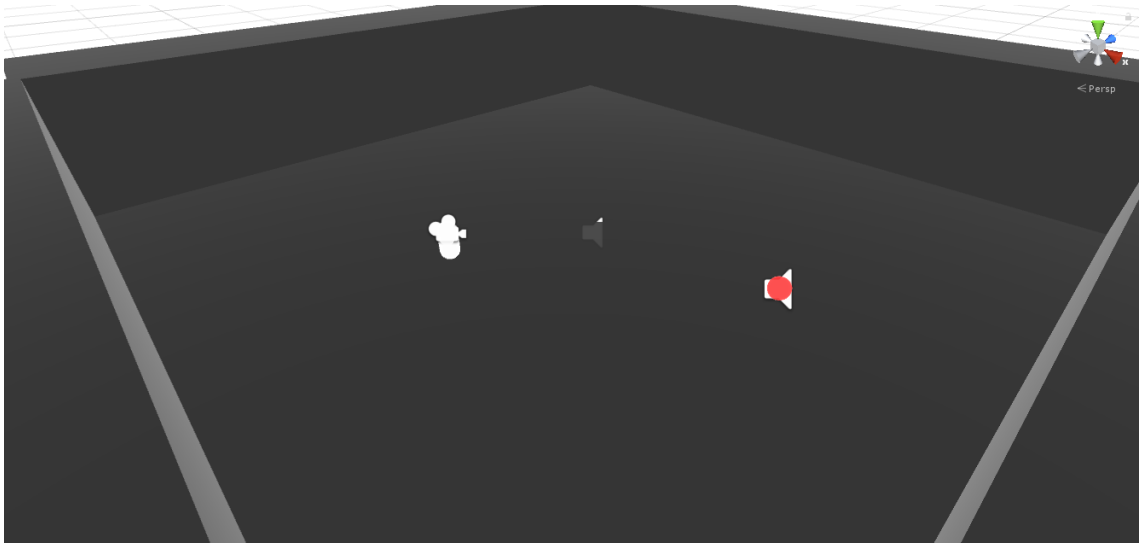


FIGURE 8.1: A top-down, 3-D rendering of the virtual game world, illustrating the position of the player avatar (the white camera) relative to a sound source (the red sphere).

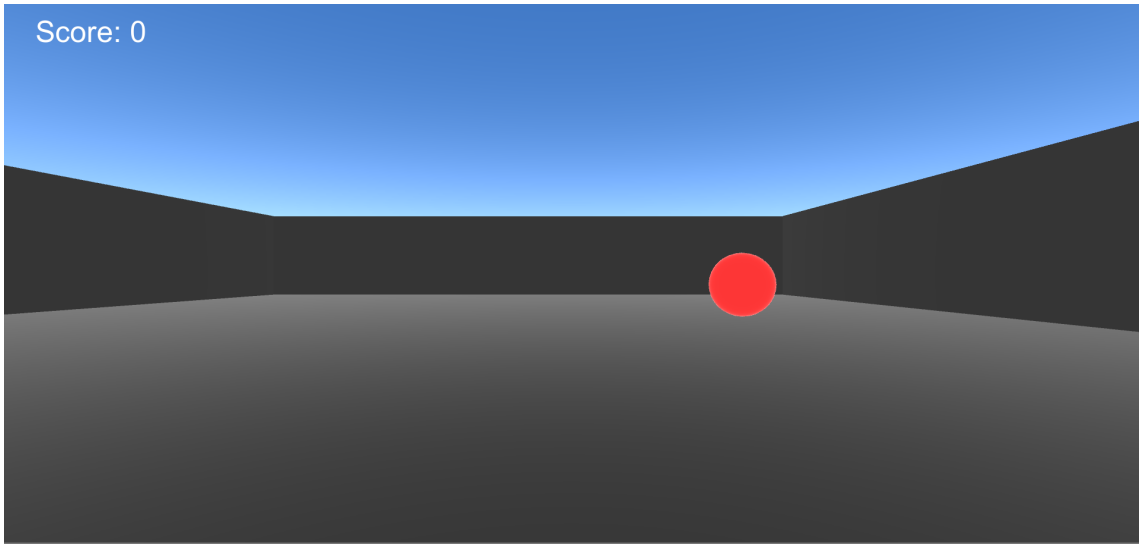


FIGURE 8.2: First-person game view from the perspective of the player avatar, with the localisation score displayed in the top left. The red sphere is the Unity game object used to represent the position of the sound source in the game world. The sphere cannot be seen whilst playing the game.

least 10 metres away from the player, within the room's boundaries. Upon triggering this event, an on-screen value depicting the player's score increased by one. A top-down interpretation is illustrated in Figure 8.3, where position A represents the current position of the sound-source and position B is the new position. If the 'x' button was pressed and the player was not within the radius of the sound source then the current position was maintained with no increase in score. A count-down timer set to 2 minutes 30 seconds was also implemented. The timer was not displayed to players and once it reached 0,

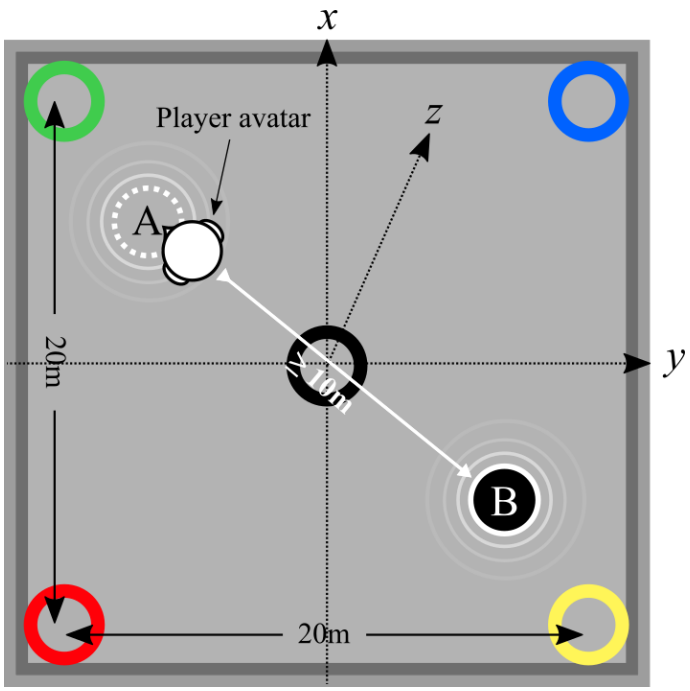


FIGURE 8.3: A conceptual illustration of a player correctly locating the sound source in its current position (A) by entering its radius and pressing 'x' on the gamepad. The sound source then moves to a new random position (B) at least 10m away from the player's current position. The coloured rings were displayed only in the training and not during the main test.

"Game Over" was displayed to the player, along with their final score to signify the end of the game session. The game was played three times by each participant. The game used throughout this experiment can be found on the attached data CD, following the index in Appendix D.4.

8.4.2 Loudspeaker Game Audio Rendering

The exact loudspeaker placements/angles for each listening condition are given in Figure 8.4, with the 7.0 arrangement conforming to the angles suggested in ITU-R BS: 775 [86].

For the loudspeaker conditions, game audio was rendered separately to the main game using the *Spatialisateur* (Spat~) object library for Max/MSP provided by IRCAM [108]. Headphone rendering was done within Unity. Communications between Unity and Max/MSP were achieved using the User Datagram Protocol (UDP) [249]. The player avatar's x, y and z coordinates in the game world were packed and transmitted over UDP on every frame update of the game. This ensured the Max/MSP patch was synced to the game systems and visuals. The x, y and z coordinates of the sound source relative

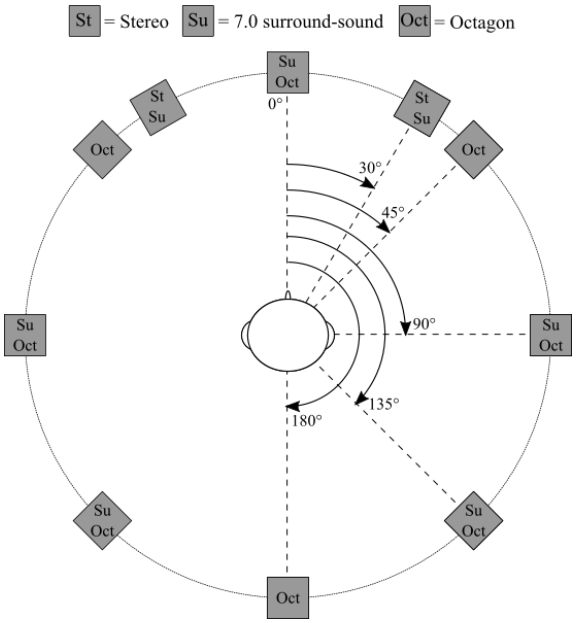


FIGURE 8.4: Loudspeaker angles used for all three listening conditions. Angles are symmetrical to the left and right of a front-facing listener.

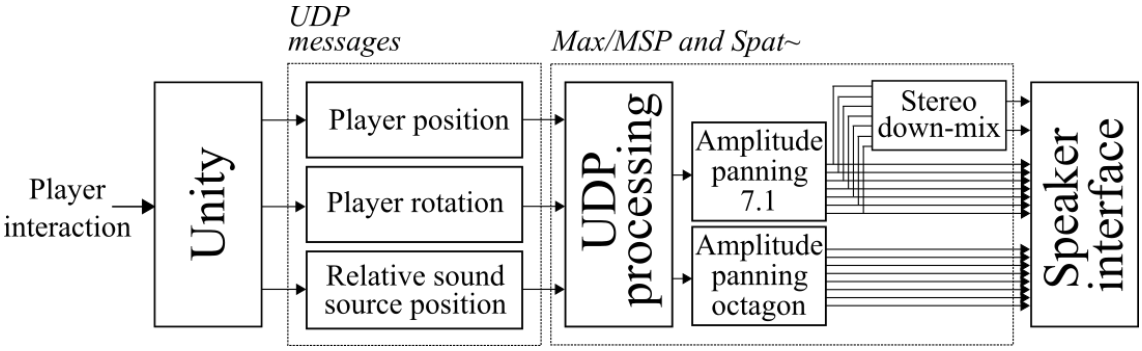


FIGURE 8.5: Outline of data flow from Unity to Max/MSP to the loudspeaker interface. Coordinates from Unity were sent to the Max/MSP patch via UDP. UDP messages were then processed to pan a sound source at different positions for the three listening conditions.

to the player were sent to Max/MSP in the same way. A diagram of the data flow from Unity to Max/MSP is given in Figure 8.5. The sound source used for localisation was a sine tone at a frequency of 440Hz repeating every half a second with an attack time of 5 milliseconds to give a hard onset. A short delay was also applied to the tone, giving a sonar-like effect. An ascending sequence of tones was played to the player upon every correct localisation to give some auditory feedback as to their success, in-line with the increase in score. If the player was incorrect, a descending sequence was played. It was decided other effects commonly found in video games, like music, ambiance and footsteps, would not be included for this test, so as to not confuse the listener.

Sound Spatialisation

In this work, pairwise panning was implemented for the 7.1 and octagonal loudspeaker configurations using a '*spat.pan*$\sim$' Max/MSP object. This method of panning retains consistency with the studies discussed in the Introduction. The '*spat.pan*$\sim$' object takes a sound source (in this case the 440Hz repeating sine tone) as its input, and pans it according to x, y and z coordinates around the pre-defined loudspeaker layout. The x, y and z coordinates used for panning correspond to the relative position of the sound source to the player, as transmitted from Unity via UDP. The number of loudspeakers and their placement around the listening area are defined for the panners as follows:

7.1 surround-sound: $0°, 30°, 90°, 135°, -135°, -90°, -30°$

Octagon: $0°, 45°, 90°, 135°, 180°, -135°, -90°, -45°$

These angles for the 7.0 condition were defined such that they conformed to the ITU-R BS: 775 for surround-sound listening [86]. The octagonal array was arranged in the same configuration used by Martin et al. [91], inclusive of a front centre loudspeaker at $0°$ relative to a forward-facing listener. Adjacent loudspeakers were positioned equidistantly with an angle of $45°$ between them. The angles used for both conditions are reflected in Figure 8.4 by the 7.0 surround-sound and Octagon labelled loudspeakers. The output for the stereo condition was generated by down-mixing the 7.0 audio using the equations suggested in [86], as presented in Section 3.3.2. The process attenuates the centre and remaining surround channels, then combines these signals with the front left/right channels, allowing listeners to experience surround-sound material at the expense of fuller spatialisation. Because all three of these conditions were rendered to loudspeakers, no additional BRIR processing was necessary.

Distance Attenuation

Since the player was able to move around the game world freely, it was necessary to include distance attenuation in the audio rendering. This made the sound appear louder as the player moved towards its source and quieter as they moved away. This was achieved by taking the inverse square of the relative distance (in metres) between the sound source and the player. This can be expressed in decibels (dB) using:

$$10 \log_{10} \left(\frac{1}{d^2} \right) \tag{8.1}$$

where d is the distance between the sound source and the listener. The same distance attenuation was used across the three conditions in order to keep changes in amplitude consistent. The amplitude of the sound remained constant as the player stayed within the radius of the sound source. This was done after informal testing, as it was found that otherwise, the sound would only ever reach maximum amplitude if the player was stood directly in the centre of the sound source position.

8.4.3 Unity Headphone Plug-in

Games authored in Unity are able to output two-channel stereo and so it is possible to generate virtualised versions of the loudspeaker conditions for playback over headphones, without the need to handle audio separately in Max/MSP, as was the case with the loudspeaker rendering in Section 8.4.2. For the current study, rendering over headphones was done using a Unity plug-in provided by DTS, a company specialising in surround-sound technologies. The plug-in itself is proprietary, therefore it is not possible to go into the exact details of its workings, however, an overview of the basic functionality is given here. For Unity, the plug-in is comprised of specialised spatial audio 'listener' and 'sound source' objects. The audio listener was attached to the player avatar, acting as a pair of ears in the virtual game world. When sounds are 'heard' by the listener, they are then rendered to a number of virtual audio channels, the arrangement of which depends on the number of virtual loudspeakers required for the VHT system. For this study seven channels were required for 7.0 surround-sound and eight for the octagonal arrangement, positioned as in Figure 8.4. The virtual channels

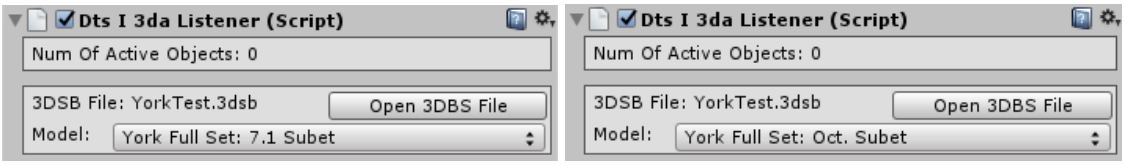


FIGURE 8.6: Screen-shots of the DTS listener object as seen in the Unity user-interface. This object is attached to the player avatar (see the white camera in Figure 8.1) and set to either 7.1 or Oct, depending on which VHT renderer is needed. For the purposes of this study the LFE channel signified by the '.1' in the surround-sound set is ignored.

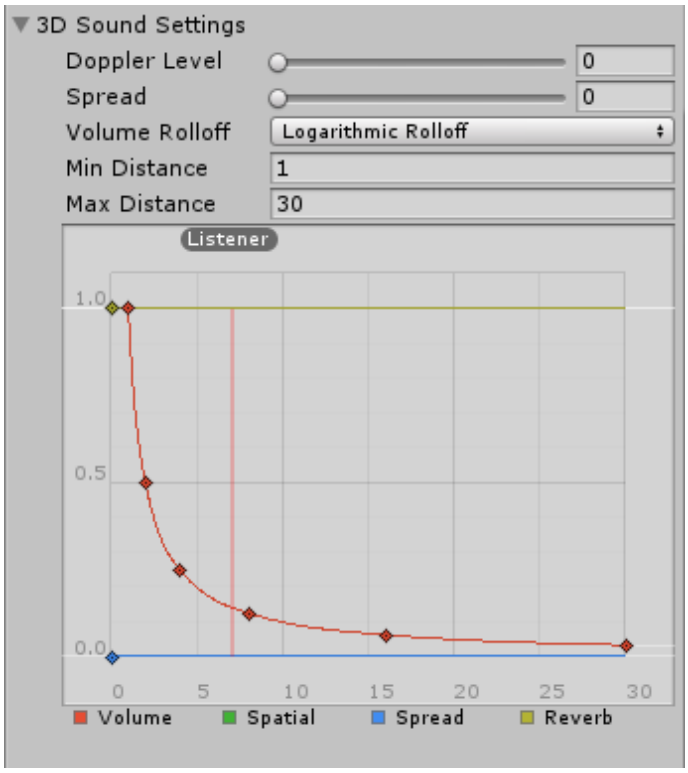


FIGURE 8.7: The logarithmic roll-off settings used in Unity to simulate distance attenuation as the player moved away from or towards the sound source.

were then convolved with a set of BRIR measurements, which are discussed in Section 8.4.3, producing a two-channel stereo output for playback over headphones.

The DTS sound source object emits audio in the game world and is attached to the same spherical game object used in the loudspeaker condition, introduced in Section 8.4.1. This means that the same random positioning of the sound source is implemented in the headphone rendering as in the loudspeaker rendering case, providing consistency between the two sets of scenarios. Sound source distance is handled by the in-built Unity roll-off function (see Figure 8.7), set to logarithmically attenuate the loudness of the source as the relative distance between it and the player avatar increases. This is essentially equivalent to the inverse-square relationship used in the Max/MSP audio engine, see Section 8.4.2. The default Unity audio listener automatically down-mixes game audio to regular stereo, therefore the DTS listener was replaced with this for the stereo experimental condition.

Gathering of BRIR

The virtual audio channels in Unity were convolved with a generic set of binaural room impulse response (BRIR) measurements corresponding to the loudspeaker positions of the 7.0 surround-sound and octagonal array. In Chapter 7, generic HRTF measurements were used for the VHT condition. The lack of room reflections may have caused the virtual loudspeakers to not be fully externalised, resulting in little perceptual difference between that and the stereo down-mix. In gathering BRIR measurements, the acoustic qualities of the room are also captured, which is known to improve externalisation in headphone based listening systems [56, 71, 73, 250]. BRIR measurements were used in the current study to investigate if improved externalisation of the virtual loudspeakers is helpful for interactive sound source localisation.

Impulse responses were recorded using a pair of G.R.A.S. 40AF microphones, built into the ear canals of a KEMAR dummy head and torso, see Figure 8.8. This is a standard technique for acquiring generic BRIR measurements. The placement of the microphones captures sound as it would transmit from the outer ear, through the ear canal to the inner ear. Measurements were taken for each loudspeaker in the 7.0 surround-sound and octagonal arrangements, giving a corresponding BRIR pair, one for each ear. Exactly the same positions and distance from the listener for the loudspeaker conditions were used, as introduced in Section 8.4.2 and Figure 8.4. The swept sine method was used as audio stimuli output from each loudspeaker. This had a duration of 2.5 seconds, sweeping up in frequency from 10Hz to 22.05kHz, at a sample rate of 44.1kHz. A proprietary file type was then generated for the DTS plug-in from each pair of recorded swept sines. These files were used for convolving the measured BRIR with the respective in-game audio channel by the DTS plug-in.

8.4.4 Training Session

Before the formal test began, participants were asked to complete a training session based on a simplified version of the game, allowing them to become familiar with the control scheme. The training version of the game took place in the same game environment, with the addition of 5 coloured rings placed at the center and each corner. These are represented by the coloured rings on the top-down game concept in Figure 8.3. During the training, the sound source would only ever appear at one of these

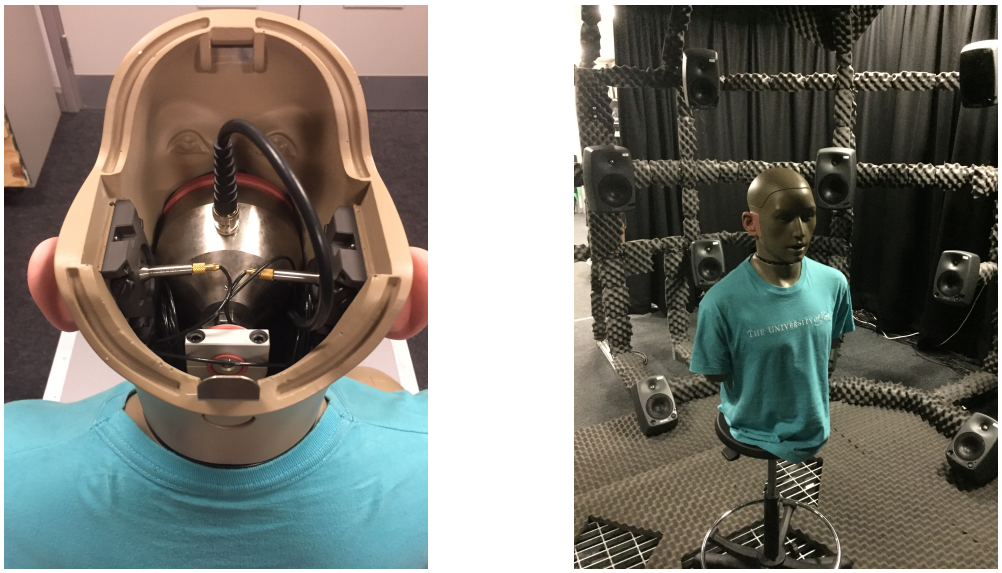


FIGURE 8.8: The KEMAR dummy head and torso used to gather generic BRIR measurements for the virtual 7.0 surround-sound and octagonal loudspeaker arrays. The picture of the left shows the placement of the two microphones in the KEMAR head.

pre-defined locations. Participants were asked to move the in-game avatar to each of these locations and press the gamepad's 'x' button if they believed that to be the origin of the sound source. Once each of the pre-defined sound sources had been found, the coloured rings were removed, and participants were asked to find the sound sources again, without a visual cue. Training was done in mono to eliminate the possible learning effect due to playing the game in an experimental condition more than once. For group A the mono channel was output from the loudspeaker positioned at 0° in Figure 8.4, whilst for group B this was output evenly from both headphone capsules. The distance attenuation was preserved, allowing participants to familiarise themselves with amplitude changes as they moved closer to and further away from the sound source. The training session was not timed and only finished once a participant had found each of the 5 sound sources twice.

8.4.5 Apparatus

For group A, 10 Genelec 8040a loudspeakers were arranged as shown in Figure 8.4, 1.5 metres from the central listening position. Those intended for 7.0 surround-sound listening conformed to ITU-R BS: 755 [86]. The Unity game and Max/MSP patch were run from the same Windows PC. Participants interacted with the game using a standard Playstation 4 gamepad connected to the PC via USB. Loudspeakers were driven by a

MOTU PCI-424 soundcard. Visuals were presented using an Optoma HD200X projector, projecting onto an acoustically transparent screen. Loudspeakers positioned at 0° and $\pm 30^\circ$ were located behind the screen.

For group B, game audio was played back using a pair of Beyerdynamic DT 990 Pro open-back headphones driven by a Sound Devices USBPre 2 Portable High-Resolution Audio Interface, connected to the PC via USB. Again, the Unity game was run from a Windows PC and controlled using a Playstation 4 gamepad. Game visuals were presented using a 24 inch HD PC monitor.

8.5 Analysis of Results

This section presents the results from statistical analysis of the player performance during the experiment for the loudspeaker and headphone based listening conditions. Player scores (i.e. the number of correct localisations) were compared between pairs of the three listening conditions within each group. Relationships between participants' success at the game and their preference for a listening condition are also given. Generally, the analysis between the loudspeaker conditions are given first, followed by equivalent comparisons between the three headphone-based conditions. The analysis is separated in this way because some of the participants in group A (loudspeakers) differed to those from group B (headphones). All statistical analysis was performed using the statistics and machine learning toolbox in MATLAB. The data and Matlab scripts for analysis can be found on the attached data CD, following the index in Appendix D.5.

The purpose of the test was to determine whether the difference between the multichannel rendering systems used for game audio playback had an impact on the number of correct localisations, and whether this impacted the player experience, as inferred from a preference score. This gave three null hypotheses to consider within each main participant grouping:

1. There is no statistically significant difference in the number of correct localisations between pairs of listening conditions.

2. There is no statistically significant difference in the route directness (i.e. the deviation between the path taken and the shortest route) between pairs of listening conditions.
3. There is no statistically significant difference in preference between pairs of listening conditions.

8.5.1 Player scores

Within groups A and B, player scores for the three listening conditions were first checked for normal distribution using a Kolmogorov-Smirnov goodness-of-fit test. Scores were found to be non-normally distributed (non-parametric), therefore sign-tests were used to check for significances between pairs of conditions within each group, as suggested by [242]. Scores were standardised before analysis due to the overall differences in scores between participants. For example one participant scored in the range of 20 to 25 during their three game sessions whilst another obtained overall much lower scores between 6 and 9. Analysing these raw values might result in the data being skewed, hence the necessity for standardisation. This was done by subtracting a participant's mean score from their three individual condition scores. This ensured the relative distances between a player's own scores would be preserved whilst being centred around 0. The output from the sign-tests for group A are presented in Table 8.2 and in Table 8.3 for group B. A value of 1 in the column labeled h of Table 8.2 and 8.3 signifies a rejection of the null hypothesis at the $p < 0.05$ significance level.

Group A

For the loudspeaker conditions used for game audio playback by group A, analysis showed there was a statistically significant difference in scores between stereo and 7.0 surround-sound as well as between 7.0 surround-sound and the octagon. Upon viewing the boxplot given in Figure 8.9 it can be seen that participants achieved higher localisation scores in the 7.0 surround-sound condition compared to both stereo and the octagonal array.

The null hypothesis could not be rejected for the comparison between the loudspeaker based stereo and octagon conditions, showing there was no statistically significant difference in player scores between. This is reflected by the boxplot in Figure 8.9, where

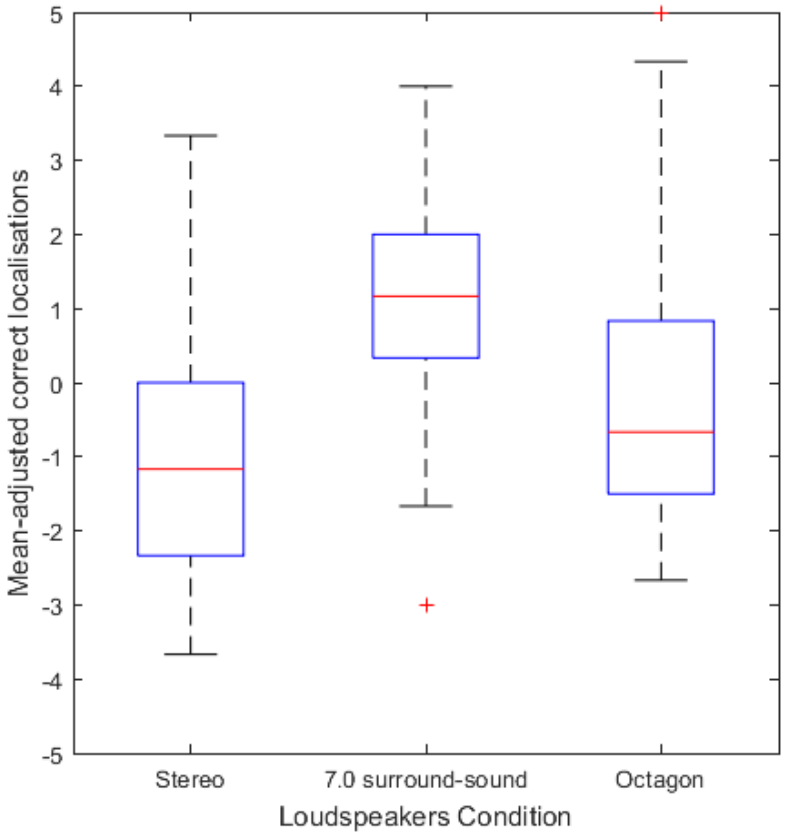


FIGURE 8.9: Mean-adjusted distribution of player scores for the three loudspeaker based listening conditions: stereo, 7.0 surround-sound and octagon within participant group A. Analysis suggests the highest scores were achieved during the 7.0 surround-sound loudspeaker condition.

it can be seen a similar range in values is spanned by the stereo and octagon plots. The result implies that participant performance neither improved nor worsened between the two conditions, in that the number of correct localisations was similar.

Group B

The comparisons between 7.0 surround-sound and the other conditions however were not the same for the headphone based rendering used by Group B. In reference to the h column in Table 8.3, it was not possible to reject the null hypothesis for any comparison between the three conditions. There was no statistically significant difference between the scores obtained for the three headphone based listening conditions, showing that the experimental condition had no impact on player performance. There is a wide

Conditions compared		Median		T	p	z	r	h
Stereo	7.0	-1.167	1.167	3	0.001	-3.198	0.067	1
Stereo	Octagon	-1.167	-0.667	7	0.136	-1.492	0.031	0
7.0	Octagon	1.167	-0.667	16	0.029	2.182	0.046	1

TABLE 8.2: Sign-test output for the mean-adjusted player scores obtained by group A. T is the signed-rank and p is the significance value. The z value is used to determine the significance value (p) and the effect size (r). A value of 1 in the h column signifies a rejection of the null hypothesis.

distribution of scores for each of the conditions, as illustrated by the boxplot in Figure 8.10, providing no clear trend in the scores obtained by players.

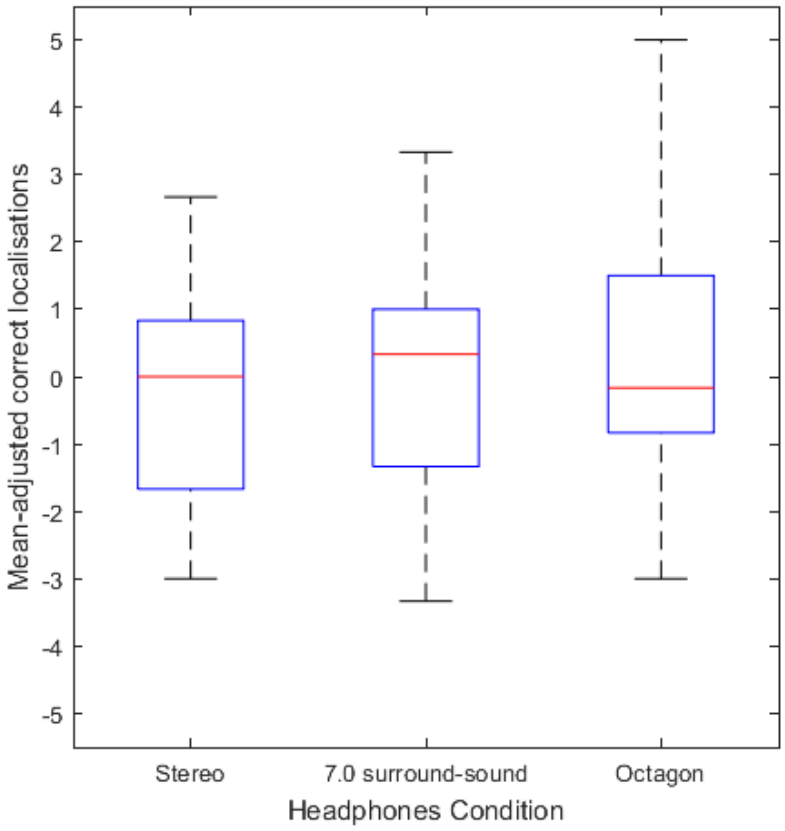


FIGURE 8.10: Mean-adjusted distribution of player scores for the headphone conditions used by participant group B. There is no clear difference between the obtained localisation scores, as reflected by the analysis in Table 8.3.

8.5.2 Route Directness Index

Whilst playing the game, the virtual route taken by each participant in each experimental condition was recorded along with the player scores and output/stored as

Conditions compared		Median		T	p	z	r	h
Stereo	7.0	0	0.333	11	0.823	0.224	0.005	0
Stereo	Octagon	0	-0.167	10	0.831	-0.213	0.004	0
7.0	Octagon	0.333	-0.167	11	1	0	0	0

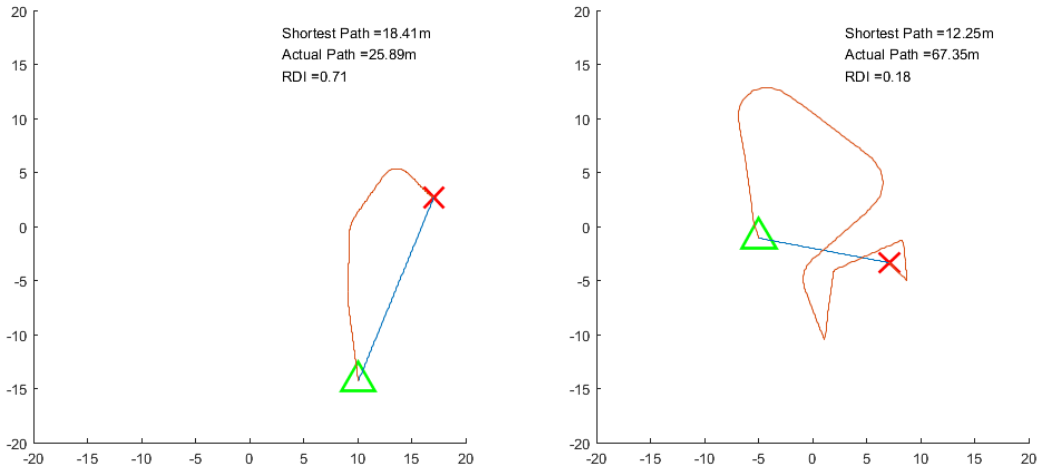
TABLE 8.3: Sign-test output for the mean-adjusted player scores obtained by group B. T is the signed-rank and p is the significance value. The z value is used to determine the significance value (p) and the effect size (r). A value of 1 in the h column signifies a rejection of the null hypothesis.

a separate text file. This was done so that the actual path taken in the virtual world could be compared with the shortest possible path between the starting position of the in-game player and that of the sound source, giving an indication as to how direct the navigation to each sound source was. The route directness index (RDI) quantifies this comparison as a numerical value between 0 and 1, where a higher value signifies greater similarity to the shortest possible route [251]. This is defined as:

$$RDI = \frac{e}{r} \tag{8.2}$$

where e is the Euclidean distance between the starting position of the in-game player avatar and the new sound source position, giving the shortest possible path between the two locations. In Figure 8.11, examples of these shortest pathways are represented by the blue lines. r is the measured distance of the actual path taken by the player from the previous sound source position to the new one. The units for distance in Unity are equivalent to meters, therefore the total length of the in-game player path and shortest path, output from the game are given in meters. For analysis, the mean RDI for each participant was compared by sign-tests, between pairs of experimental conditions in the same way as the player scores.

For both groups A and B, the RDI values were not significantly different between any of the listening conditions and thus the null hypothesis could not be rejected for either. This shows that none of the listening conditions had an effect on the virtual path taken within the game world. Referring to Figures 8.12 and 8.13, the boxplots are relatively similar in shape, with very little difference between the median lines, which also indicates that the RDI values were relatively similar across all three conditions for the two groups. It is interesting to note the similarity of the boxplot median lines between all three of



(a) High RDI example

(b) Low RDI example

FIGURE 8.11: Example plots of the actual path taken by the player (orange line) compared to the shortest path (blue line). The RDI for plot A is higher because the in-game distance travelled by the participant is similar to the shortest path. The axes represent the dimensions of the virtual game world in meters.

the headphone conditions and the stereo and octagon loudspeaker conditions given in Figure 8.12, implying consistently low levels of directness towards the sound source in these cases.

Conditions compared		Median		T	p	z	r	h
Stereo	7.0	0.408	0.526	7	0.066	-1.837	0.038	0
Stereo	Octagon	0.408	0.450	13	0.838	0.204	0.004	0
7.0	Octagon	0.526	0.450	15	0.307	1.021	0.021	0

TABLE 8.4: Sign-test output for the player RDI in the loudspeaker conditions. T is the signed-rank and p is the significance value. The z value is used to determine the significance value (p) and the effect size (r). A value of 1 in the h column signifies a rejection of the null hypothesis.

8.5.3 Player Preference

Once participants had played the game using all three listening conditions, they stated on a questionnaire which of the three was preferred, and were also encouraged to provide comments regarding their decision. For group A, 7.0 surround-sound was the most preferred of the three, as chosen by 70.8% of participants. Tables 8.6 and 8.7 show the percentage of highest player scores attained in each condition, alongside the corresponding percentage of overall preference, for loudspeakers and headphones

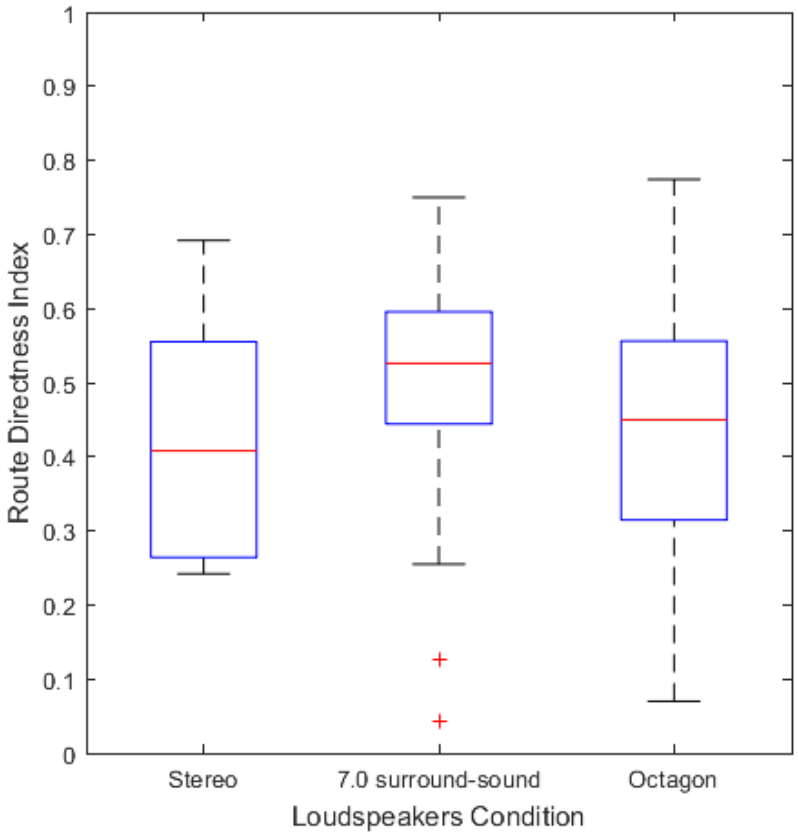


FIGURE 8.12: Distribution of the route directness index (RDI) for the three loudspeaker conditions: stereo, 7.0 surround-sound and octagonal array. Analysis suggests that the most direct routes to the sound sources were taken during the 7.0 surround-sound condition.

respectively. For the loudspeaker conditions, 60.4% of the highest scores were obtained in the 7.0 surround-sound condition, which is also most preferred by 70.8% of players. Both stereo and the octagonal array were preferred by significantly fewer participants. The preference scores given by group A for each condition are illustrated by the bar chart in Figure 8.14a. Conversely, there is no clear majority in the number of preference ratings given for the headphone conditions by group B, in that all three conditions were preferred by less than 50% of the total participants (see Figure 8.14b). This indicates that preference ratings between players were not as consistent when using headphones, implying similar experiences were had among different participants across the three headphone conditions.

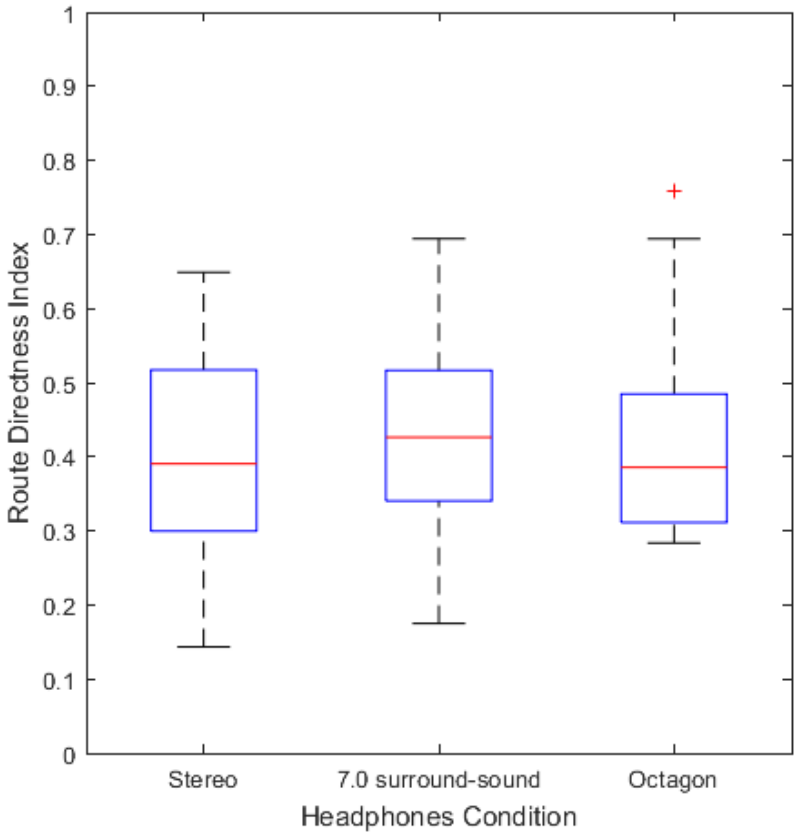


FIGURE 8.13: Distribution of the RDI values for the three headphone based experimental conditions. The similarity between the three plots suggests that the routes taken by participants were neither more nor less direct.

8.6 Discussion

When considering only the results obtained from the loudspeaker conditions, a higher majority of players clearly had greater success in the game when listening to audio over a 7.0 surround-sound loudspeaker array, in comparison to audio presented using either stereo or an octagonal array. The number of correct localisations was consistently higher, where 60% of participants received their highest scores when using surround-sound. When considering the sign-test comparison between stereo and 7.0 surround-sound, the effect size ($r = 0.514$) also signifies that listening condition had a large effect on player scores. This is higher than the moderate effect size observed between the 7.0 and octagon conditions ($r = 0.292$). For loudspeakers, this suggests scores achieved in the 7.0 condition were consistently higher in comparison to stereo than when compared to the octagonal array.

Conditions compared		Median		T	p	z	r	h
Stereo	7.0	0.391	0.426	9	0.307	-1.021	0.021	0
Stereo	Octagon	0.391	0.386	9	0.307	-1.021	0.021	0
7.0	Octagon	0.426	0.386	13	0.838	0.204	0.004	0

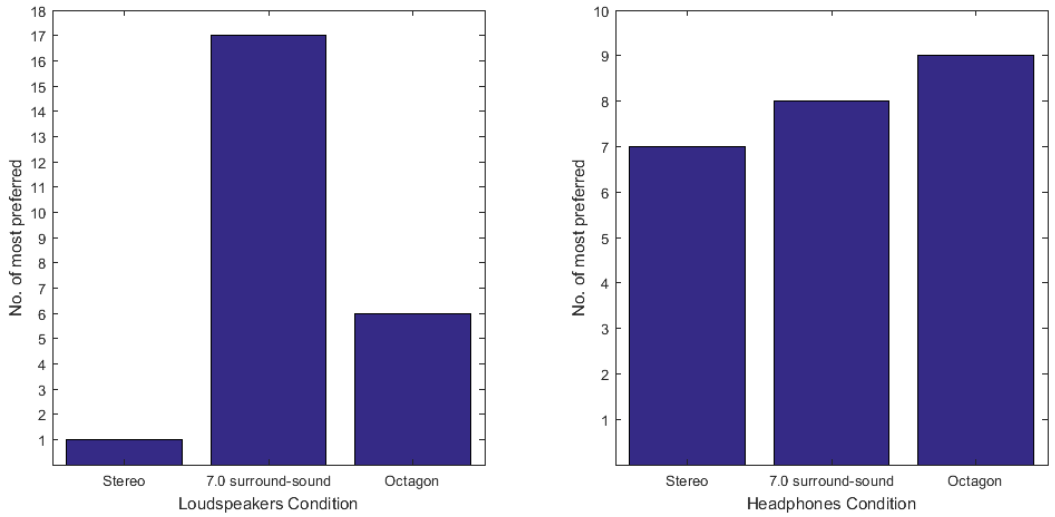
TABLE 8.5: Sign-test output for the comparison of RDI between the headphone listening conditions. The value of 0 in column *h* suggests that the null hypothesis cannot be rejected for any of the comparisons.

7.0 was also the most preferred of the three listening conditions, with comments from some participants (specifically participants 5, 14 and 19 in Appendix C.4.1) suggesting this was due to it being the condition in which the highest scores were achieved. It was expected that loudspeaker 7.0 would outperform stereo due to the increased number of channels available in the system, and from these results it can be said that players did benefit from using a listening array with rear and side channels. In Chapter 6 it was observed that 7.1 surround-sound had higher perceptual spatial quality than regular two-channel stereo, which were both assessed here. In relation to those results, this implies that there is a positive relationship between the perceived spatial quality of a multichannel listening system and the in-game performance of a player. This provides evidence for the idea that 7.1 surround-sound game audio with high perpetual spatial quality offers an advantage to video game players.

Unexpectedly, similar trends were not observed for the physical octagon loudspeaker array. Based on work by Theile and Plenge [90] and Martin et. al [91], it was expected that the localisation of sound sources, especially those positioned laterally and to the rear of the listener, would be easiest when listening over an octagonal array of loudspeakers. However, the more consistent and stable phantom imaging that can be achieved using such a system seems to have had little impact on the results obtained in

Condition	% highest score	% most preferred
Stereo	12.5%	4.2%
7.1 surround-sound	60.4%	70.8%
Octagon	27.1%	25.0%

TABLE 8.6: The percentage of highest scores achieved, alongside the percentage of preference ratings, for each loudspeaker condition, as experienced by group A. The majority of high scores were attained when listening over 7.0 surround-sound, suggesting this is why it was most preferred.



(a) (b)

FIGURE 8.14: Preference ratings given for the loudspeaker (a) listening conditions and headphone (b) conditions. 7.0 surround-sound was preferred by the majority of participants who were exposed to the loudspeakers. There is no clear majority in the number of preference ratings given for the headphone conditions.

Condition	% highest score	% most preferred
Stereo	29.2%	29.2%
7.1 surround-sound	33.3%	33.3%
Octagon	37.5%	37.5%

TABLE 8.7: The percentage of highest scores, alongside the percentage of preference ratings, for each headphone condition, experience by group B. There is no majority in preference or highest scores for any of the three conditions.

this experiment. Visuals were presented to the player using a stationary screen, therefore players were only ever required to look forwards. For this reason it may have been that those loudspeakers located directly infront of the listener's forward facing position were of most use in the localisation task. The front left and right loudspeakers of the octagonal array were spaced wider than the $\pm 30^{\circ}$ used in the 7.0 arrangement. Although a centrally placed loudspeaker was used in both the 7.0 and octagon conditions, those at the front left and right were spaced wider in the latter. The increased resolution generated by the narrower angles between the left, right and centre loudspeakers in 7.0 surround-sound may have been more helpful than consistent imaging from all directions. Also, the directionality that can be achieved with a 7.0 array, although not perfect, would at least allow a listener to gain a good sense of a

sound source's specific direction. It may therefore be the case that once a player had positioned their in-game avatar such that the sound was perceived to emanate at some point straight ahead, triangulating its specific location was then easiest using the increased frontal resolution produced by the narrower angles in the 7.0 condition. Some comments from participants provide further evidence to this, where it was stated on multiple occasions that it was easiest to triangulate/focus on the sound source in the 7.0 surround-sound condition. Participant 6 noted that sound sources were 'easiest to triangulate' using loudspeaker 7.0 surround-sound, and other comments from participants 1, 2, 7, 8, 10, 12, 13, 16 and 24 reflect this, whilst also stating that localisation was quicker and easier, and also that the resolution of the condition was felt to be higher (see Appendix C.4.1). This, however, was not reflected in the analysis of player directness, where the RDI for loudspeaker 7.0 was not significantly different to any other loudspeaker condition, suggesting participants did not take a more direct route to the sound source.

Regarding the three headphone conditions, there was a similar trend in the results to those presented in Chapter 7. According to the statistical analysis, neither of the virtual loudspeaker conditions outperformed a non VHT stereo down-mix. This implies that for the presented localisation task, players did not benefit from additional playback channels when they were virtualised for headphone listening, in the same way as they would when using a physical loudspeaker array. As in Chapter 7, this brings into question the perceptual differences between VHT systems and stereo down-mixes (which do not employ binaural processing) intended for game audio headphone playback. Going by the subjective quality ratings given in Chapter 7, there was an expectation that this may be the case since the spatial attribute *localisation accuracy* was perceived to be similar across headphone 7.0 and stereo. One aim throughout this thesis has been to investigate whether a more convenient method for listening to multichannel game audio, in the form of a virtualising headphone system, has any impact on the player experience. As discussed in Chapter 5, there seems to be a belief in the non-academic gaming community that headphone based multichannel audio does influence the way in which games are played and experienced. It may be that the experimental stimuli used throughout this thesis have been unsuitable for the purposes of testing such systems. In future work it might be worth focusing on games, or

particular in-game mechanics and systems, that real game players believe to benefit from multichannel audio. Such content could be identified through an interview process and/or by searching relevant gaming related on-line forums, reviews and articles.

The use of generic BRIR measurements, rather than HRTFs, for the headphone conditions also seems to have had little impact on how well a player could localise a sound. Some players (namely participants 1, 2, 3, 4, 9 and 24 - see Appendix C.4.2) commented on the altered timbre of the emitted sound due to the additional process involved in BRIR rendering, causing them to prefer the stereo down-mix, even if their performance was better in another condition. Non-individualised BRIR measurements, as used in the surround-sound and octagon headphone conditions, have been shown to cause timbral changes, and also decrease localisation accuracy for some listeners [252, 253]. As suggested in Chapter 7 it might be beneficial to employ individualised measurements to mitigate these problems. However, it is the belief of the author that it is important to consider headphone based listening systems that are equivalent to those actually available to video game players commercially. It is unlikely that individual measurements will be viable for the vast majority of gamers, due to the fact that they generally need to be obtained in controlled laboratory conditions. It is probable that generic HRTF/BRIR measurements, usually based on a dummy head (such as the KEMAR used in this study), are used in most virtual surround-sound products, it is therefore likely that users will experience similar undesirable changes in timbre.

For the loudspeaker conditions, the analysis of preference scores suggested that there was a preference for the condition in which the game was found to be easiest, which in the majority of cases was 7.0 surround-sound. Although the stereo condition contributed to 12.5% of the highest scores, some participants who achieved those scores stated in the accompanying questionnaire that, perceptually, sounds were easier to localise in 7.0 surround-sound, hence it was more preferred. This may explain the minor discrepancy between the score and preference percentages for the stereo and surround-sound conditions. The percentage of highest scores and preference ratings were equal for the headphone (Group B) conditions, which implies that non of the headphone conditions were preferred. For some participants, their comments (see Appendix C.4.2) suggest that there was a noticeable change in timbre for the surround-sound and octagon conditions which had an influence on their given

preference. Participants 1, 2, 3, 4, 9 and 24 all made negative statements regarding the timbre of the sound in the VHT 7.0 and VHT octagon conditions. The stereo down-mix, which has no additional BRIR processing, was generally perceived to give a clearer overall sound. The change in timbre in the VHT 7.0 and VHT octagon conditions may have therefore distracted from the game session, causing the player to have a negative impression.

It is also important to note that in comparison to commercial games, the game used in this experiment was a relatively simple example. Generally, modern games include more in-depth sound design often making more complete use of surround channels, and visual effects that work together in forming the entire game experience. It would therefore be of interest to determine whether the results obtained from this experiment could be replicated using a more complex game task, inclusive of more 'true-to-life' game systems. This would provide clarity as to whether the results from this study were dependent on the stimulus used. However, this would require extensive knowledge of professional game design and a potential team of developers, and is therefore out of the scope for this current research. In comparison to the work presented in Chapters 6 and 7, however, the use of a custom game environment was found to allow for far more control over experimental variables, and is therefore recommended for such studies.

8.7 Summary

This chapter has presented an experiment designed to determine whether enhanced spatial audio feedback has an influence on how well a player performs in a video game. Player performance was quantified by how many correct localisations of a randomly positioned sound source were achieved within a time limit of 2 minutes 30 seconds, in a custom game environment authored using the Unity game engine. This was compared between down-mixed stereo, 7.0 surround-sound and an octagonal array of loudspeakers, and equivalent headphone renderings. Loudspeaker rendering was achieved using pairwise panning in a separate audio engine created in Max/MSP, whilst headphone virtualisation was done using a Unity plug-in provided by DTS. For the headphone system, generic BRIR measurements were gathered using a KEMAR dummy head for the loudspeaker positions of a 7.0 surround-sound and octagonal loudspeaker

array. This ensured the loudspeaker and headphone conditions were equivalent.

Results suggest that by using 7.0 surround-sound played back over loudspeakers, player performance was improved, in that significantly higher localisation scores were achieved in comparison to the other experimental conditions. Loudspeaker 7.0 was also consistently the most preferred by participants. Based on preference being an indicator of QoE, it can be assumed that the participants' experience was improved as a result of surround-sound in that it made the game easier to play. Results suggest that the location of the sound source was less well defined in all other experimental conditions, making the game altogether more difficult. The game session will have been made more difficult when the sound source was harder to find. This may explain the high preference for loudspeaker surround-sound where the highest majority of high scores were obtained. For the headphone conditions, the BRIR processing did not offer any advantage over a stereo down-mix. This brings into question the effectiveness of virtual headphone based surround-sound systems for video game audio playback, unfortunately suggesting that it provides no clear advantage. However, the author admits that the use of non-individual BRIR measurements and the simplicity of the game used as experimental stimuli might have been inappropriate for the localisation task.

Chapter 9

Summary and Conclusions

Before the main conclusions are made, summaries of the main results from Chapters 6, 7 and 8 are given on the following pages for reference. Table 9.1 is also given in order to make clear which of the listening systems considered in the experimental work were classified as surround-sound. The table is intended to be used in reference to the chapter summaries, as well as the restatement of the overall hypothesis given in the main Conclusions section.

Listening Condition	Chapters	No. of Discrete Audio Channels	No. of Loudspeakers (real or virtual)	Defined as surround-sound?
Mono	N/A	1	1	No
Big Mono	6	1	7	No
Stereo	6, 7, 8	2	2	No
Big Stereo	6	2	7	No
7.1 surround-sound	6, 8	7 + LFE	7 + Subwoofer	Yes
VHT 7.1	7, 8	7	7 (virtual)	Yes
Octagon	8	8	8	Yes
VHT Octagon	8	8	8 (virtual)	Yes

TABLE 9.1: A list of the listening conditions considered in each experiment, along with the number of audio channels and number of loudspeakers used to output those channels. The table also shows whether each listening condition was classed as surround-sound for reference with the hypothesis.

9.1 Chapter 6 - Perceived Spatial Quality and Player Preferences

This chapter presented a listening test designed to find if a listener believed the spatial sound quality of a 7.1 surround-sound listening system was higher than that of a stereo or mono system, whilst engaged in playing a video game. The test was also designed to

find out if the 7.1 surround-sound version of the video game audio was the most preferred. The results from an initial pilot study showed that 7.1 surround-sound was perceived to have higher spatial sound quality than stereo and was also most preferred. For the pilot study, the stereo soundtrack was output from a single right and a left loudspeaker, as would be the case in a traditional arrangement. For the main experiment, the mono and stereo listening conditions were modified such that the audio channels were output from all the available loudspeakers of a 7.1 surround-sound configuration. These were referred to as Big Mono (BMo) and Big Stereo (BSt) respectively. The results from the main experiment showed that both 7.1 surround-sound and BSt were preferred over BMo, and on a whole were perceived to have higher spatial sound quality. However, the results also showed that 7.1 surround-sound and BSt were similarly preferred and were also perceived to have similarly high spatial sound quality. The similarities between 7.1 surround-sound and BSt were unexpected, although the results did suggest that high spatial sound quality was preferable whilst playing a video game.

9.2 Chapter 7 - Headphone-Based Audio Rendering and Player Preferences

Chapter 7 presented a listening test with similar methods to those used in Chapter 6, but instead compared a VHT rendering of 7.0 surround-sound video game audio with an equivalent down-mix to stereo, both played back over headphones. Both headphone renders were done using a custom system designed in Max/MSP using Ambisonic theory and HRTF measurements gathered from a dummy head. The results showed that there was no perpetual difference in the spatial quality between the two listening conditions and neither was preferred. However, both the VHT 7.0 surround-sound rendering and the down-mix to stereo were perceived to have high spatial sound quality, again suggesting that this is desirable in a gaming context.

9.3 Chapter 8 - The Impact of Multichannel Audio on Player Performance

The final experiment presented in Chapter 8 investigated how the performance of a video game player might improve as a result of them using a surround-sound audio system. The player was asked to locate an object in a custom game environment by only listening for the sound it emitted. The success of the player was compared between stereo, 7.0 surround-sound and octagonal listening systems. Half of the participants experienced the three listening conditions over loudspeakers and the other half experienced them over headphones. Results showed that participants performed best at the game when listening to a 7.0 surround-sound version of the audio played back over a physical loudspeaker array. The 7.0 surround-sound loudspeaker array was also the most preferred. For the participants using headphones, the statistical analysis showed that there was no statistically significant difference between the listening conditions for either player performance or preference. This suggested that the difference between the listening conditions was more noticeable when the audio was played back over loudspeakers than when it was played back over headphones, which is also reflected by the results gathered in Chapters 6 and 7.

The results across all three chapters showed that high spatial sound quality was desirable regardless of the type of audio system, and player performance was improved when using a surround-sound loudspeaker arrangement. The results varied depending on the rendering format used, and there was also a more apparent difference between listening conditions played back over loudspeakers than those played back over headphones.

9.4 Conclusions

The overall hypothesis considered throughout this thesis has been the following:

> *The implementation of surround-sound in an interactive video game environment, rendered either over loudspeakers or headphones, will have a positive impact on a player's gaming experience in comparison to stereo or mono*

The overall conclusions drawn from the three listening tests presented in Chapters 6, 7 and 8 do suggest that surround-sound game audio has some influence on the player

experience, supporting the initial hypothesis for this thesis. It was shown that perceptually high spatial sound quality is preferable, although in most cases stereo and 7.1 surround-sound were rated similarly highly, which was unexpected. Notably, player performance was shown to improve as a result of using loudspeaker 7.0 surround-sound in a video game designed around an interactive localisation task. In this particular case, participants were able to more consistently localise sound sources in the loudspeaker 7.0 surround-sound condition suggesting that the task was made easier. In considering the results for the loudspeaker conditions presented in Chapters 6 and 8 respectively, there is an implication that there is a positive relationship between the perceived spatial quality of a multichannel listening system and the in-game performance of a player. From this it can be theorised that those video game players who have access to loudspeaker surround-sound systems will have a potential advantage over those who do not, although to quantify this would require further study. This opens up some interesting avenues for continued research into the field, where tests could be designed to more formally assess the differences in player performance between different listening systems. These might involve competitive tasks between multiple game players to assess whether a player using surround-sound can more easily out-perform another using stereo.

For all of the headphone based conditions presented in both Chapters 7 and 8, results suggested that VHT surround-sound systems do not out-perform more standard stereo down-mixes. This is the case for the subjective attribute ratings in that the spatial quality was perceived to be similarly high across all conditions, and also for the player performance scores, which did not improve whilst using virtual surround-sound. VHT systems were not preferred in either experiment, suggesting that players found it hard to distinguish them from the stereo down-mix. As suggested in Chapter 7 it might be beneficial to employ individualised measurements to mitigate the potential perceptual problems caused by using HRTF/BRIR measurements taken using a dummy head. However, it is the belief of the author that it is important to consider headphone based listening systems that are equivalent to those actually available to video game players commercially. It is unlikely that individual measurements will be viable for the vast majority of gamers, due to the fact that they need to be taken in controlled laboratory conditions. It is probable that generic HRTF/BRIR measurements, usually based on a

dummy head, are used in virtual surround-sound products and therefore it is likely users will experience undesirable changes in timbre [252, 253]. This is unfortunate, as headphone systems offer a far more convenient alternative for surround-sound listening, providing a potential for their relative benefits compared to loudspeaker based playback. This has been one of the motivations for this thesis, where there seems to be a belief in the non-academic gaming community that headphone based surround-sound audio does influence the way in which games are played and experienced [11–16]. It may be that the experimental stimuli used throughout this thesis has been unsuitable for the purposes of testing such systems, although it was the intention in the experimental work to use as true to 'real-life' examples as possible. In future work it would be of interest to focus on games, or particular in-game mechanics and systems, that real game players believe to benefit from surround-sound, rather than those chosen by the author alone.

Lessons can also be learned by reflecting on the different test methodologies used throughout this thesis. From the first two experiments, it became apparent that gathering subjective data, such as spatial audio attribute ratings, may not have been suitable. The attention needed to play a game for the required amount of time will have made it difficult for participants to fully focus on the required audio rating tasks. Also, using a pre-existing game became a problem where the absence of control of the underlying systems made it difficult to ensure multiple participants would have similar exposures to the stimuli. By simplifying the game task and relying on objectively gathered data, the methodology employed in Chapter 8 made it much easier to interpret any differences between experimental conditions. It is therefore recommended that test designs using objectively measurable metrics, with a clear task for the player, are more appropriate for any experimentation involving video games.

In conclusion the results presented in this thesis have shown that perceptually high spatial sound quality is desirable for video game audio playback, suggesting that ongoing research and development into surround-sound audio systems for gaming is necessary for improving the overall player quality of experience. Results also indicated that player performance was improved when audio was played back over a 7.0 surround-sound loudspeaker system as opposed to stereo or an octagonal array. This is a key finding which suggests that spatial audio can contribute significantly to tactical

player advantages. As such, this finding can impact on the future development of video games, showing audio can be used as another tool for developers to influence player decision making. Additionally, the experiment in Chapter 8 found that the use of surround-sound changed the way in which players interacted with the virtual space providing a stimulus for further research into audio-led navigation in video games and interactive media. Another noteworthy finding was that the effect of different headphone solutions was not as clear as the differences between the equivalent loudspeaker solutions. This may be due to a lack of personalised HRTFs for the headphone renderings which is one of the major challenges in current research for spatialised headphone playback in interactive media applications. In the wider context of consumer multichannel systems for game audio, a balance has yet to be struck between the practicalities of headphone presentation and the more consistent spatial imaging provided by loudspeakers. This thesis has shown that exploring these options for game audio playback is a worthwhile course for future development, since player experience and performance were improved as a result of using surround-sound.

9.5 Further Work

The conclusions drawn from the work presented in this thesis offer some interesting opportunities concerning possible further work into the field of multichannel audio in video games. The author believes there is significant creative potential in how video game related listening tests can be designed and implemented. This section will point out and list some ideas that would benefit from further investigation.

Compare generic and individualised HRTF and/or BRIR measurements in a gaming context:

- Individualised measurements might offer a greater perceptual difference between VHT systems and a stereo down-mix.
- BRIR measurements for the participants from the experiment presented in Chapter 8 have already been obtained. This will provide a basis for comparison to the KEMAR measurements already employed.

Generate more complex and representative video game environments for testing:

- The game presented in Chapter 8 is a relatively simple example, and does not represent the potential of most modern, commercially available games.
- Improvements could be made by adding more sound effects that react to player input, including music and improving the visual quality. Steps would however have to be taken to ensure that these improved elements do not distract from the task influenced by the main research question under consideration.
- The author believes using custom game environments is beneficial as more control can be had over what the player can and cannot do. This allows for increased repeatability between participants, as well as between experimental sessions.

Formally investigate competitive advantages between players:

- Results from Chapter 8 suggest that loudspeaker 7.0 surround-sound made the game easier for some participants, providing a tactical advantage. It would be interesting to test this in a more competitive environment between multiple players simultaneously.
- The experiment might involve training a group of video game players to perform equally at a game in a controlled condition, such as mono. Players could then compete, each using a different audio playback condition, to explore how these systems might offer an advantage. For example, it is anticipated, although as yet unproven, that a player using 7.1 surround-sound might more easily beat another using stereo.
- This would require in-depth knowledge of game design, involving systems to allow two or more players to interact simultaneously, and potentially build into the game more complex environments, decision making and tasks to be completed.

Explore multichannel audio in the context of less typical game genres:

- As stated in Chapter 4, it was difficult to find examples of effective surround-sound implementation outside of some very specific genres (for example: shooters, horror and action).
- The use of spatial sound in some other genres might offer some creative avenues for game design, rather than in just enhancing the general audiovisual aesthetic.

Spatial audio could be integrated more into the design of the rules and mechanics controlling the game world.

Formally analyse the correlation between spatial attribute ratings and preference:

- For the results presented in Chapters 6 and 7, the relationship between the spatial quality ratings given and the preference scores were only inferred from the individual analysis of those results.
- It would be more formal to correlate the quality and preference scores, however it was felt doing this retrospectively would impact the design of subsequent experimental work.
- Correlating the results in this way would more definitively show which spatial attributes contributed significantly to the overall preference score, as is the recommendation for the *analytic* quality assessment methods introduced in Chapter 5.
- New subjective listening tests could be devised, expanding on those given in Chapters 6 and 7, to gather information on more spatial attributes, and relate those attributes to the overall experience of the player.

9.6 Final Remarks

The purpose of this thesis has been to explore how surround-sound audio might be used in video games based on the idea that they are well suited to this kind of audio presentation. Sounds can react dynamically to player input, offering clear advantages, in terms of aesthetics and tactics. The results in this thesis have demonstrated that surround-sound can be preferable in a gaming context whilst also offering an improvement to player performance. However, this is not to say that game experiences are lacking without surround-sound, as it is clearly the case that the vast majority of gamers have their expectations fulfilled, regardless of whether or not they own a surround-sound system. In fact, through browsing support forums for various games and game playing systems, it becomes apparent that there is some confusion in the wider gaming community concerning the purposes of multichannel audio, and the options that are available (as discussed in Chapter 5. Nevertheless, in the past decade

there has been an increasing trend for hardware and software developers (especially those in the AAA title market) to implement enhanced multichannel audio, with 7.1 surround-sound being standard for any games developed for the PlayStation 4 or Xbox One. It is therefore useful to consider the more creative ways in which this ever evolving technology can be used in a way to drive in-game decisions, as suggested in Chapter 8, rather than just as a tool to enhance subjective audio quality. There are already examples of games in which audio is used as an influencing factor on gameplay decisions, such as *Papa Sangre 2* [168] and *Hellblade: Senua's Sacrifice* [169], and the advent of virtual reality is also an important milestone, with companies such as Google and Sony researching further into interactive and immersive audio playback. The conclusions drawn from this thesis have only scratched the surface on the advantages of immersive audio in video games and there are clearly many creative avenues to be explored if the role of multichannel game audio as an influence on player quality of experience is to be more fully understood.

Appendix A

Perceived Spatial Quality and Player Preferences Experiment Pack

This appendix provides the experiment pack given to participants for the experiment presented in chapter 6. The layout is as follows:

1. **Experiment Information Sheet** - a document outlining the experiment and what is expected of the participant.
2. **Consent Form** - a form requiring the signature from each participant to state that they agree with and have read the information sheet.
3. **Spatial Attributes Reference** - the list of spatial attributes with descriptors used by participants to rate the sound quality of each listening condition. *Sound source definition* and *Stability* were omitted from the analysis in chapter 6.
4. **Event Time Line** - this document provides participants with a time line of events from the introductory sequence of *The Last of Us: Remastered*.
5. **Control Scheme** - a diagram of the control scheme for *The Last of Us: Remastered*.
6. **Questionnaire** - the questionnaire used by participants to rate spatial quality and state preferences.
7. **Demographic Information** - a document for participants to fill in their demographic information.
8. **Comments** - a list of participant comments from the experiment presented in chapter 6.

A.1 Experiment Information Sheet

Experiment Information Sheet

You are being invited to take part in an experiment investigating the question: 'Do video game players prefer playing video games in a spatial listening environment?' Please take the time to read this information sheet to understand what this study hopes to achieve and what you will be asked to do, before you decide whether or not you wish to take part.

The purpose of this experiment is to find if a more spatially capable sound system, for example multichannel surround sound, is preferable to video game players in comparison to more common loudspeaker system setups, such as mono or stereo, due to the fact it can more appropriately translate spatial information to the listener.
You do not have to take part in this study if you do not want to. If you wish to continue please read the rest of this information sheet and sign the consent form. After reading this information sheet and signing consent you may still withdraw from the experiment at any point without giving reason.

You will be asked to subjectively assess different listening environments by rating separate spatial audio attributes believed to account for the overall spatial quality of a listening system. These attributes will be explained to you in detail by the experimenter, so don't worry if initially you find them hard to understand. You will then be asked to rate which out of the listening environments you are personally exposed to is most preferred. Questionnaires and a list of attributes with descriptors will be available to you from now for the duration of the experiment.

The experiment will require you to play a segment of the Playstation 4 title *The Last of Us: Remastered*. This game has a PEGI (Pan European Game Information) 18 rating as it contains strong language and extreme violence which some participants may find offensive and distressing. If this does not sound like the type of game you would like to play, due to its explicit content, then it is advised that you do not take part in this study. If you are not a fan of violent or scary films, it is also advised you do not take part in this study. The game has been chosen as it is relatively easy to play and has a clear, linear path for you to follow; therefore no prior game experience is necessary.

Any personal information given will be obtainable only by the experimenter and project supervisor and will be stored on a password protected University computer as well as a central filestore. Names will not be used in obtaining subject responses so anonymity will be preserved.

Results obtained from this study will be used in the experimenter's MSc by Research thesis and in any related publications by the experimenter. If you wish to know the results of this test or have any questions after the experiment has taken place please contact the experimenter via email (jrj504@york.ac.uk).

Thank you for taking the time to read this information sheet and I hope you will participate.

A.2 Consent Form

Consent Form

1. I can confirm that I have read and fully understood the information sheet. ☐
2. I acknowledge the explicit content of this study and am willing to participate at my own discretion. ☐
3. The experimenter has made it clear I can leave the experiment at any time without giving reason. ☐
4. I can confirm I am of the age 18 or over. ☐

_____ _____

Date Signature

A.3 Spatial Attributes Reference

Spatial Attributes Reference

This table can be used as a reference for the spatial attributes you will be asked to rate. Please take the time to read through the table and ensure you have a clear understanding of each attribute. If the meaning of a particular attribute is unclear please feel free to ask the experimenter for a more detailed description.

Depth	Refers to the perceptual front-back distance of an auditory scene. The perceived distance between different sound sources will create an overall soundscape receding from the listener. A scene with a good sense of Depth will help to create a sense of perspective in the soundscape, suitable to the virtual setting.
Distance	Refers to the perceived distance of a sound source. A high rating would be given if there is good agreement between a sound sources perceived distance and the position of its respective in-game object (the virtual object producing the sound).
Envelopment of reverberation	The extent to which the player feels spatially surrounded by the reverberant energy in the virtual space. A high rating will include a level of envelopment, according to the acoustical properties of the environment.
Sound source localisation	A sound source is suitably positioned in the virtual space in relation to its in-game object (the object that is supposed to be producing the sound) and the player and output to the appropriate speaker(s) – you can identify from what direction a sound source is coming from.
Sound source definition	Multiple sound sources heard at the same time can be clearly identified and separated from one another.
Stability	Sound sources that you can hear stay in their intended positions relative to the position of the player camera (listener).
Source width	Refers to how large a space a sound source is perceived to occupy in a horizontal direction. Source width can be perceived by the appropriate mapping of a sound to multiple speakers at the same time. A large sound source, close to the player, might be heard from two or three speakers, giving a sense of its size.

A.4 Event Time Line

Event time line

This document has been created to give you a rough guide as to what you should expect in the play-through and roughly how long it will take. You may consult this document at any point during the experiment.

00:00 – For the first part of the game you will play as Sarah – a 12-year old girl. After getting out of bed, explore the house and interact with objects, like doors and a mobile phone, using 'Δ' when prompted. You will eventually end up downstairs in a kitchen area.
Sounds to note: A television displaying the news; an explosion in the distance; police car sirens; mobile phone; whimpering dog.

02:44 – You will hear shouting from outside. Joel (Sarah's father) will enter through the patio doors and shoot an infected neighbour.
Sounds to note: Infected neighbour banging on the patio window/door.

03:43 – Tommy (Joel's brother) pulls up outside the house and Joel and Sarah enter the car. You are free to move around in the back seat of the car to get a better idea of the events unfolding around you.
Sounds to note: Farm house on fire; police car sirens; family standing on the side of the road.

06:04 – The car is forced to stop due to a traffic jam. An infected hospital patient attacks a bystander in front of you and another runs towards the car. Tommy reverses the car and attempts to find an alternate route through town. Again you are free to move around in the back seat of the car.
Sounds to note: Infected patient banging on car window.

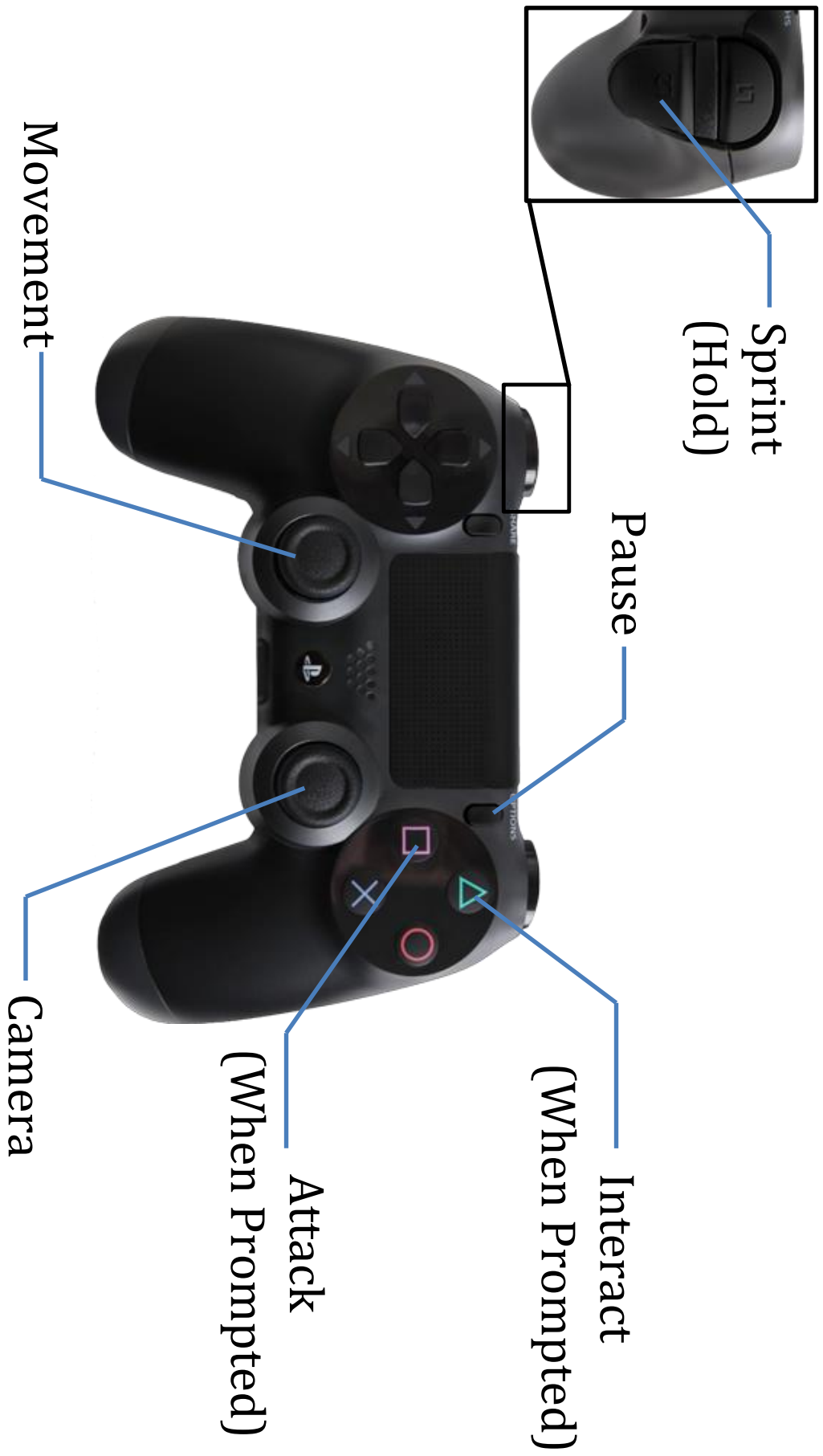
06:57 – Whilst navigating through the town an out of control lorry slams into the side of the car, flipping it over. You are now playing as Joel and need to escape the upturned car. Repeatedly press '□' when prompted to break the front window and escape from the car.
Sounds to note: Impact of lorry hitting car.

08:05 – Joel picks up Sarah and begins to navigate through the town on foot carrying her. You will need to find a safe path through the town and avoid any contact with infected citizens. Push the left analogue stick forward to make Joel move forward and use the right analogue stick to rotate the camera – which is how you can view or look around the scene. Remember - you cannot defend yourself with a young girl in your arms so your only option is to run!
Sounds to note: Turmoil of the town; gas station exploding; car crash; cars on fire.

09:06 – Follow Tommy down the alleyway in front of the flaming cars. Halfway down the alley you will be confronted by an infected citizen. Repeatedly tap '□' to push the attacker away, and then let Tommy finish him off. Run into the bar at the end of the alley. Tommy agrees to stay behind to hold off any more infected people, allowing you (Joel) to escape with Sarah.
Sounds to note: The ambience of the alley compared to the more open town environment.

09:58 – Leave the bar and escape into the outskirts of town. You will be chased by a number of infected civilians, do not stop running! If you are caught you will have to start from outside the bar again. Eventually you will come across a soldier who shoots the pursuers. A short cinematic scene will play marking the end of the play-through.
Sounds to note: The running and breathing of the infected civilians as they chase you.

A.5 Control Scheme



A.6 Questionnaire

After each play-through please grade each individual spatial attribute on the scale given below by placing a cross in the applicable box.

Play-through #1 - Listening environment 'A' spatial sound quality

	BAD	POOR	FAIR	GOOD	EXCELLENT
DEPTH					
DISTANCE					
SOUND SOURCE LOCALISATION					
SOUND SOURCE DEFINITON					
STABILITY					
REVERBERATION					
SOURCE WIDTH					

Play-through #2 - Listening Environment 'B' spatial sound quality

	BAD	POOR	FAIR	GOOD	EXCELLENT
DEPTH					
DISTANCE					
SOUND SOURCE LOCALISATION					
SOUND SOURCE DEFINITON					
STABILITY					
REVERBERATION					
SOURCE WIDTH					

Once you have completed both play-throughs please indicate which you preferred the most and to what extent by placing a cross in the appropriate box below.

STRONG PREFERENCE FOR 'B'	PREFERENCE FOR 'B'	SLIGHT PREFERENCE FOR 'B'	NO PREFERENCE	SLIGHT PREFERENCE FOR 'A'	PREFERENCE FOR 'A'	STRONG PREFERENCE FOR 'A'

A.7 Demographic Information

Participant No.:_____

Please fill in this form before the experiment begins. All personal information will be kept anonymous.

1. Age: _____

2. Gender: _____

3. Do you have any previous listening test experience? YES/NO

4. On average, how often do you play video games?

Daily	_____
Several times a week	_____
Several times a month	_____
Several times a year	_____
Never	_____

5. What is your preferred gaming platform?

Console (e.g PS4/Xbox 360/Nintendo Wii)	_____
PC/Mac	_____
Mobile device/tablet	_____
Handheld (e.g Nintendo 3DS/PS Vita)	_____
None	_____

A.8 Participant comments

- **Participant 5** - 'The second [Big Stereo] one was much better, particularly with elevation. The helicopter was an example - I could really hear it going overhead the second [Big Stereo] time. '
- **Participant 6** - 'Sometimes sound sources in A [Big Stereo] seemed quite excessively spread out, also sometimes sources go further to L [left] and R [right] than on-screen location.'
- **Participant 15** - 'Felt B [Big Stereo] was easier to localise [than 7.1 surround-sound].'
- **Participant 19** - 'Potentially would have preferred to rate attributes after both play-throughs (easier comparison/point of reference).

Appendix B

Headphone Based Audio Rendering and Player Preferences Experiment Pack

This appendix provides the experiment pack given to participants for the experiment presented in chapter 7. The layout is as follows:

1. **Experiment Information Sheet and Consent Form** - a document outlining the experiment and what is expected of the participant. The document also acts as the consent form.
2. **Demographic Information** - a document for participants to fill in their demographic information.
3. **Event Time Line** - this document provides participants with a time line of events from the introductory sequence of *The Last of Us: Remastered*.
4. **Spatial Attributes Reference** - the list of spatial attributes with descriptors used by participants to rate the sound quality of each listening condition.
5. **Questionnaire** - the questionnaire used by participants to rate spatial quality and state preferences.
6. **Comments** - a list of participant comments from the experiment presented in chapter 7.

B.1 Experiment Information Sheet and Consent Form

Experiment Information

You are being invited to take part in an experiment investigating spatial audio implementation in video game environments. Please take the time to read this information sheet to understand what you will be asked to do, before deciding whether or not you wish to take part.

The experiment will require you to play two segments of the Playstation 4 title *The Last of Us: Remastered*. This game has a PEGI (Pan European Game Information) rating of 18 as it contains strong language and extreme violence. If you find this kind of explicit content offensive or distressing then it is advised that you do not take part in this study. The game's audio will be played to you over a pair of headphones in both game sessions. Please do not adjust the volume during the test.

You will be asked to give a subjective rating for a number of attributes relating to spatial audio quality after each game session. Descriptors for these attributes will be available to you for the duration of the test. Finally, you will be asked to rate which of the two game sessions you preferred the most and to what extent.

Any personal information given will be obtainable only by the experimenter and project supervisor and will be stored on a password protected University computer as well as a central filestore. Names will not be used in obtaining subject responses so anonymity will be preserved.

Results obtained from this study will be used in the experimenters PhD thesis and in any related publications by the experimenter. If you wish to know the results of this test or have any questions after the experiment has taken place please contact the experimenter via email (jrj504@york.ac.uk).

You do not have to take part in this study if you do not want to. If you wish to continue please sign the attached consent form and return it to the experimenter. After giving consent you may still withdraw from the experiment at any point without giving reason.

Thank you for taking the time to read this document and I hope you will participate!

1. I can confirm that I have read and fully understood the information sheet.
2. I acknowledge the explicit content to be used in the study and am willing to participate at my own discretion.
3. The experimenter has made it clear I can leave the experiment at any time without giving reason.
4. I can confirm that I am of the age 18 or over.

_____ _____

Signature Date

B.2 Demographic Information

Participant No.:

Please fill in this form before the experiment begins. All personal information will be kept anonymous.

1. Age:
2. Gender:
3. Do you have any previous listening test experience? **YES**/**NO**
4. Have you played *The Last of Us* or *The Last of Us: Remastered*? **YES**/**NO**
5. On average, how often do you play video games?
   - Daily □
   - Several times a week □
   - Several times a month □
   - Several times a year □
   - Never □

B.3 Event Time Line

Event Time Line

This document has been created to give you a rough guide as to what you should expect during the play-through. You may consult this document at any point during the experiment.

00:00 - For the first part of the game you will play as Sarah - a 12-year old girl. After getting out of the bed, explore the house and interact with objects, like doors and a mobile phone, by pressing △ when prompted. You will eventually end up downstairs in a kitchen area.
Sounds to note: A television set; an explosion in the distance; police car sirens; mobile phone; whimpering dog

02:44 - You will hear shouting from outside. Joel (Sarah's father) will enter through the patio doors and shoot an infected neighbour.
Sounds to note: Infected neighbour banging on the patio window/door.

03:43 - Tommy (Joel's brother) will pull up outside the house and Joel and Sarah enter the car. You are free to move around the back seat of the car to get a better idea of the events unfolding around you.
Sounds to note: Farm house on fire; police car sirens; family standing on the side of the road.

06:04 - The car is forced to stop due to traffic. An infected hospital patient will attack a bystander and another will run towards the car. Tommy reverses the car and attempts to find another route through town.
Sounds to note: Infected patients banging on car window.

06:57 - Whilst navigating through town an out of control lorry will slam into the side of the car, flipping it over. You will now play as Joel and need to escape the upturned car. Repeatedly press □ when prompted to break the front window and escape from the car.
Sounds to note: Impact of lorry hitting car

08:05 - Joel lifts Sarah and begins to navigate through the town. You will need to find a safe path through the town and avoid contact with infected citizens. Push the **left analogue stick** forward to make Joel move forward and use the **right analogue stick** to rotate the camera. Remember - you cannot defend yourself with a young girl in your arms so your only option is to run!
Sounds to note: Turmoil of the town; gas station explosion; car crash; flaming cars

09:06 - Follow Tommy down the alley. Halfway down the alley you will be confronted by an infected citizen. Repeatedly tap □ to push the attacker away. Run into the bar at the end of the alley. Tommy will agree to stay behind to hold off any more attackers, allowing you to escape with Sarah.
Sounds to note: Alley ambience

09:58 - Leave the bar and escape into the outskirts of town. You will be chased by a number of infected civilians, do not stop running! Eventually you will come across a soldier who shoots the pursuers. A short cinematic scene will play marking the end of the play-through.
Sounds to note: Your pursuers

B.4 Spatial Attribute Reference

Spatial Attributes Reference

Please take the time to read through this reference sheet and ensure you have a clear understanding of each attribute. If the meaning of a particular attribute is unclear please feel free to ask the experimenter for a more detailed description.

Localisation Accuracy: Refers to how well you can identify the direction in which a sound source is originating. There should be good agreement between the visual location of an object and the sound in emits.

Distance Accuracy: Refers to the perceived distance of a sound source. There should be good agreement between a sound sources perceived distance and the position of its related in-game object.

Sense of Depth: Refers to the perceived front-back definition of the sound scene and the sound sources within it. A scene with a good sense of depth will help to create a sense of auditory perspective.

Sense of Width: Refers to the perceived left-right definition of the sound scene and the sound sources within it.

Envelopment: Refers the extent to which you as the player feels surround by the sound in the presented scene.

B.5 Questionnaire

After each play-through please grade each **spatial attribute** on the scale below.

Play-through A Spatial Quality

	Bad	Poor	Fair	Good	Excellent
Localisation Accuracy	◯	◯	◯	◯	◯
Distance Accuracy	◯	◯	◯	◯	◯
Sense of Depth	◯	◯	◯	◯	◯
Sense of Width	◯	◯	◯	◯	◯
Envelopment	◯	◯	◯	◯	◯

Play-through B Spatial Quality

	Bad	Poor	Fair	Good	Excellent
Localisation Accuracy	◯	◯	◯	◯	◯
Distance Accuracy	◯	◯	◯	◯	◯
Sense of Depth	◯	◯	◯	◯	◯
Sense of Width	◯	◯	◯	◯	◯
Envelopment	◯	◯	◯	◯	◯

Once you have completed **both** play-throughs please indicate which was most **preferred** and to what extent on the scale below.

Strong Preference for A	Preference for A	Slight Preference for A	No Preference	Slight Preference for B	Preference for A	Strong Preference for A
◯	◯	◯	◯	◯	◯	◯

This space has been left blank for any **comments** you may have.

This is the end of test.

B.6 Participant Comments

- **Participant 1** - 'Voice of playable character seemed to be coming from left - a little disconcerting. TV static in test B [VHT] seemed to be coming from 2 distinct positions when panning behind - didn't move smoothly as in test A [stereo down-mix].'
- **Participant 3** - 'It was best with the surround [VHT], also was very good with the head-tracking.'
- **Participant 17** - 'A [stereo down-mix]: Panning too hard. No sense of depth for voice acting. B [VHT]: More accurate voice location. Too mushy for distances.'
- **Participant 18** - 'Zombie voices from A [VHT] seemed to be coming from inside my head.'

Appendix C

The Impact of Multichannel Audio on Player Performance Experiment Pack

This appendix provides the experiment pack given to participants for the experiment presented in chapter 8. The layout is as follows:

1. **Experiment Information Sheet and Consent Form** - a document outlining the experiment and what is expected of the participant. BRIR measurements were only taken for participants using the headphone conditions and were not used in this thesis.
2. **Consent Form** - a form requiring the signature from each participant to state that they agree with and have read the information sheet.
3. **Questionnaire and Demographic Information** - a document for participants to fill in their demographic information and state which of the three listening conditions was preferred.
4. **Comments** - a list of participant comments from the experiment presented in chapter 8.

C.1 Experiment Information Sheet

Experiment Information

You are being invited to participate in an experiment investigating the implementation of spatial audio in video game environments. Please take the time to read this information sheet to understand what you will be asked to do, before deciding whether or not you wish to take part.

The experiment will require you to complete a task in a custom game environment designed to assess the importance of specific sound qualities in spatialiased game audio. The nature of the task will depend on the spatial sound qualities under investigation and will be made clear to you by the investigator before the game session. You will be asked to complete a short training session to become familiar with the control system and in-game task. Depending on your group allocation, the game's audio will be presented to you using either an array of loudspeakers or a pair headphones. Audio output will be set to a comfortable listening level before the test. Please **do not** adjust this level.

When the game session is complete, you will receive a score informing you of your performance. You will also be asked to fill out a short questionnaire concerning your experience of the game session and your thoughts on the audio presentation method used. You will also be asked to take part in a process that will measure how sound propagates from different directions around your head to your ears. From these measurements a set of filters, unique to you, will be generated. Once the measurements have been processed you will be asked to return at a later date to repeat the localisation test using your own personalised filters. This data will not be used in any work other than that by the investigator and will not be publicly available. The measurement process will take no more than 15 minutes and consist of the following steps:

1. You will be asked to take a seat in the loudspeaker rig and make yourself comfortable.
2. Small microphones will be placed at the entrance of your ear canals. This will be similar to wearing a pair of in-ear headphones.
3. You will then be asked to sit up straight and keep still whilst the measurement is happening,
4. The measurement will begin and you will hear sine tones, rising in frequency, from different points around the measurement rig. Try to keep as still and quiet as possible.
5. The measurement is now finished. Please await instructions from the investigator.

Any personal information given will be accessible to the investigator and project supervisor only. You will be asked to leave an email address so you can be contacted at a later date. You will only be identifiable by a unique participant number attached to this email address. This information will be stored on a password protected computer and physical copies will be kept secure in a locked office. Names will not be used in obtaining subject responses so anonymity will be preserved.

Results obtained from this study will be used in the investigators PhD thesis and in any related publications by the investigator. If you wish to be kept informed on the progress of the study or have questions after the experiment has taken place, please contact the experimenter via email (jrj504@york.ac.uk).

You do not have to take part in the study if you do not want to. If you wish to continue please sign the attached consent form and return it to the investigator. After giving consent you may withdraw from the study at any point without giving reason. If you decide to withdraw from the test any data collected will not be used and will be destroyed.

Thank you for taking the time to read this document and I hope you will participate.

C.2 Consent Form

Please place a cross (**x**) in the boxes provided.

- ☐ I can confirm that I have read and fully understood the information sheet.
- ☐ The investigator has made it clear I can leave the experiment at any time without giving reason.
- ☐ I acknowledge that I will be contacted by the investigator to return at a later date to repeat the experiment and am able to decline this offer.
- ☐ I can confirm that I am of the age 18 or over.

Signature:

Date:

Email address:

C.3 Questionnaire and Demographic Information

Participant number:

Presentation Order:

Please fill in this part of the form before the experiment begins. All personal information will be kept anonymous.

1. Age:
2. Gender:
3. Do you have any previous listening test experience? **YES/NO**
   If **yes**, please give a brief summary:

4. Are you familiar with handling a video game controller/joypad? **YES/NO**
5. Would you consider yourself a video game player? **YES/NO**
6. On average, how often do you play video games?
   - ☐ Daily
   - ☐ Several times a week
   - ☐ Several times a month
   - ☐ Several times a year
   - ☐ Never

Fill in this part of the form after you have completed the experiment.

Please indicate with a cross (**x**) which of the three game sessions you preferred:

- ☐ A (first session)
- ☐ B (second session)
- ☐ C (third session)

If applicable, please specify what influenced your preference:

Any other comments:

This is the end of the test

C.4 Participant Comments

C.4.1 Group A - Loudspeaker Playback

- **Participant 1** - 'Easier to pick out direction of sound more quickly [using 7.1 surround-sound] than the first session [stereo].'
- **Participant 2** - 'Third [7.1 surround-sound] better than first [stereo], not much difference between B [octagon] and C [7.1 surround-sound] but C seemed a bit easier.'
- **Participant 3** - 'B [stereo] slightly more spatially telling. C [octagon] was definitely the least clear, but more used to game by then.'
- **Participant 4** - 'Better spatialisation [in 7.1 surround-sound]. Did not like C [stereo]! Sounded mono.'
- **Participant 5** - 'Got the highest score [using 7.1 surround-sound], don't think I could tell much difference [perceptually]. In 2 [stereo] and 3 [7.1 surround-sound] I felt like there was more sound above me.'
- **Participant 6** - 'I found it easiest to triangulate position in B [7.1 surround-sound]. It was much harder to do so in C [stereo] so I ended up looking around more.'
- **Participant 7** - 'Felt the imaging was more reliable [in reference to 7.1 surround-sound] - surround version were far better than stereo, and thought B [7.1 surround-sound] was more accurate/close to the source.'
- **Participant 8** - '[In reference to 7.1 surround-sound] Rotation as well as translation felt like it was more helpful than in other sessions. Knowing where middle of sound was easier.'
- **Participant 9** - '[In response to octagon] more directional response when moving through the game. 3rd [octagon] best - good user feedback, wide but concise direction. 1st [7.1 surround-sound] second best - hard when moving. 2nd [stereo] worst - also hard when moving, narrow directionality.'
- **Participant 10** - 'Found it generally easier to localise the sound source in first test [7.1 surround-sound]. With the third one [stereo] it felt easier when i got more used to the spatialisation.'

- **Participant 11** - 'It [the first session (octagon)] is the most challenging session as I was not familiar with the system at the beginning and have not developed a personalised playing style for the game.'
- **Participant 12** - 'Sound was easier to locate when moving. Easier to find the centre of sound [in reference to 7.1 surround-sound].'
- **Participant 13** - 'I think it was easier to determine the localisation of the source, because I think sound had better resolution [in reference to 7.1 surround-sound].'
- **Participant 14** - 'Second session [octagon] was more difficult to locate when moving at angles rather than rotating and moving forward. Hard to distinguish between first [stereo] and third [7.1 surround-sound] session. Chose third session [7.1 surround-sound] for preference due to higher score separating the two. Third session [7.1 surround-sound] has a more nuanced feel. Reverberant qualities were better in third [7.1 surround-sound] over the first [stereo].'
- **Participant 15** - '[In reference to octagon] Easier to localise the sound. Easier once I had more practice.'
- **Participant 16** - 'Easier to locate the sound accurately [using 7.1 surround-sound]. For B [octagon] and C [stereo] it is harder to find the specific spot.'
- **Participant 17** - 'Because I had more practice in playing [in reference to choosing condition C (7.1 surround-sound) as the most preferred].'
- **Participant 18** - 'I thought A [octagon] was easier.'
- **Participant 19** - 'Did the best the second [7.1 surround-sound] time.'
- **Participant 20** - 'Easier to hear sound from left or right [using octagon].'
- **Participant 21** - 'Found it hard to locale sound on B [stereo]. On C [octagon] I felt it was easier to find but couldn't get it exactly. A [7.1 surround-sound] felt like a middle ground.'
- **Participant 22** - 'Second [octagon] clearer and easier to locate.'
- **Participant 23** - '[7.1 surround-sound] felt more natural than the other two.'

- **Participant 24** - 'A [octagon] and B [7.1 surround-sound] seemed very similar but B [7.1 surround-sound] better. C [stereo] a bit hazy lacking in azimuth.'

C.4.2 Group B - Headphone Playback

- **Participant 1** - 'C [VHT octagon] was hard - easily got lost. B [VHT 7.0] didn't sound good. Filtering was obvious [for VHT conditions].'
- **Participant 2** - 'Clearer sound [when using stereo]. Filtering was obvious [for VHT conditions].'
- **Participant 3** - 'Balance between localisation accuracy and timbre [for stereo]. Less front-back confusion in third session [VHT octagon], but filtering was obvious.'
- **Participant 4** - 'B [VHT octagon] seemed to have easier front-back localisation and more tonally even. A [VHT 7.0] sounded filtered and had poor front-back, so spent a lot of time trying to work that out. C [stereo down-mix] seemed to be harder to exactly pinpoint the sound but had good front-back.'
- **Participant 5** - 'First [VHT octagon] was difficult to tell front-back, could only really tell left-right. Third [VHT 7.0] felt more clearly localised in space than second [stereo down-mix].'
- **Participant 6** - 'It was slightly easier to hear where the sound was coming from [using stereo down-mix].'
- **Participant 7** - '[VHT Octagon] sounded better/easier to tell direction. A lot of front-back confusion on both 2nd [VHT 7.0] and 3rd [stereo down-mix sessions]. Gaps in sound between each burst because I would move and wouldn't know where the sound had gone!'
- **Participant 8** - '[In reference to VHT 7.0] Sound source location was more defined. However, some small problems with front and back.'
- **Participant 9** - 'A [VHT 7.0] and C [VHT octagon] sounded tinnier. C [stereo] hardest to locate when far away.'
- **Participant 10** - 'Better azimuth cues [using VHT Octagon], good rearward position. A [VHT 7.0] was phasey, C [stereo down-mix] was bad.'

- **Participant 11** - 'C [VHT 7.0] was a little harsh, A [VHT Octagon] was a bit too jumpy/switched left-right quickly. B [stereo down-mix] felt most natural.'
- **Participant 12** - 'C [stereo down-mix] - no externalisation and very little panning. B [VHT 7.0] - more difficult I think. Not good externalisation on any. Easier to move around until it got louder than to actually look around for sound.'
- **Participant 13** - 'Second [VHT 7.0] was hard to place L-R.'
- **Participant 14** - 'Spatialisation [using VHT 7.0] felt clearer, I found the sounds easiest to find. However, it was a close call between A [stereo down-mix] and C [VHT 7.0] as C's sounds felt smaller and quieter.'
- **Participant 15** - 'That one [VHT octagon] sounded most realistic, very 3D, easy to identify where the sound was coming from. B [stereo down-mix] was the worst one, very mono with little 3D effect. Relied on amplitude differences.'
- **Participant 16** - 'B [VHT octagon] and C [stereo] was easier to locate the sound, B [VHT octagon] was clearest in terms of direction and easier to tell when right.'
- **Participant 17** - '[VHT 7.0] caused less confusion and I got the highest score out of the 3.'
- **Participant 18** - 'Found it easiest to locate the sound [using stereo down-mix].'
- **Participant 19** - '[In reference to VHT 7.0] I felt more responsive to the noise in the space, so it gave more of a challenge. I also preferred the noise.'
- **Participant 20** - 'First [stereo down-mix] sounded disjointed/inconsistent, difficult to navigate. Second [VHT octagon] sounded a bit warbly, but easier to navigate. Third [VHT 7.0] was very easy to hear where sound was.'
- **Participant 21** - '[Using VHT octagon] sound was easier to differentiate from left or right.'
- **Participant 22** - '[In reference to VHT octagon] it was easier to locate the source. Third [stereo down-mix] was the hardest, transition between L-R not smooth.'
- **Participant 23** - 'After two attempts, it was easier, I was more familiar.'

- **Participant 24** - 'Localisation features for B [VHT 7.0], but timbre for C [stereo down-mix].'

Appendix D

Data CD Index

This index lists the contents of the attached data CD. The content is split into folders named according to the relevant thesis chapter.

D.1 Chapter 6 Data and Analysis

Contains data from experimental work, matlab scripts for analysis and a 'read me' document.

D.2 Chapter 7 Max/MSP Patch

Contains the Max/MSP patch used to render headphone listening conditions and a 'read me' document.

D.3 Chapter 7 Data and Analysis

Contains data from the experimental work, a matlab script for analysis and a 'read me' document.

D.4 Chapter 8 Game

Contains the game code used to generate experimental stimuli, the Max/MSP patch used for loudspeaker audio rendering and a 'read me' document.

D.5 Chapter 8 Data and Analysis

Contains raw data from the experimental work, matlab scripts for analysis and a 'read me' document.

Glossary

AAA video game Pronounced 'Triple A', this refers to video games with a relatively big budget, equivalent to a blockbuster film. Usually, this will mean the overall team of individuals working on the game will be large, and significant time and resources can be spent on every part of the development process. Predominant producers/developers include *EA*, *Ubisoft*, *Warner Bros.* and *Sony*.

Ambisonics A technique for encoding audio by decomposing a sound-field using spherical harmonics.

Audio listener A game object that acts as a microphone in the game world so that sound sources can be heard by the player [254]. The listener will catalogue where in the virtual environment a sound source is emitting so that the underlying audio systems can perform processes to adjust the panning and amplitude of the sound. The audio listener usually works in-sync with the viewpoint of the player so that sounds can react relative to what is also seen.

Audio middleware This is a software package, independent of the core game engine, used to control the behaviour of in-game sound effects by triggering them, cataloguing source positions and applying various manipulative effects through digital signal processing (DSP), such as reverb and equalisation (EQ). Examples of such software include *Wwise* [255] and *Fmod* [256], both of which support a wide range of multichannel rendering formats, intended for both loudspeaker and headphone playback.

Big Mono (BMo) A method whereby one discrete audio channel is duplicated and output over multiple loudspeakers.

Big Stereo (BSt) A method whereby two discrete audio channels are duplicated and played back routed to more than two loudspeakers.

Console generation This refers to the period of time in which similar video game console hardware is simultaneously developed and released by competing companies. A new console generation will begin whenever a new piece of gaming hardware is released and is considered to be a substantial upgrade to that of its predecessors.

Cut-scene Generally, a pre-rendered segment of a game used as a means to present the player with information, such as instructions or narrative elements. Cut-scenes can be thought of as small films in-between sections of gameplay, where full player interaction is usually suspended.

Game engine A framework of code and computer scripts for creating video games, usually bundled as a complete software package. The game engine is responsible for handling the game's visual rendering, physics systems, underlying rules/mechanics, human interfacing (i.e. real-world interaction and control), artificial intelligence (AI) and basic audio. Open-source game engines include *Unity 5* [248] and *Unreal Engine 4* [257] which have free licenses for personal use. Some developers have their own propriety 'in-house' engines not available to the public, often with specialist functionalities, such as *Luminous Studio* [258] used by Square Enix and CD Projekt Red's *REDengine* [259].

Game object Conceptually, objects refer to the game's building blocks. They act as containers for all the systems and code required to construct anything needed to make the game operate as intended, such as walls, characters, weapons or on-screen text [260].

Indie video game Shorthand for 'Independent', this refers to video games that are usually funded independently of a core game publisher, with a much smaller team of developers. These games are similar to independently funded films. More recently, video games have been delivered digitally over services such as *Steam*, increasing the number of indie games available to the public.

Low Frequency Effects (LFE) Channel A discrete audio channel used in surround-sound systems to separate lower frequency sound effects. This channel is often routed to a sub-woofer, rather than a full-range loudspeaker.

Mono A method for presenting audio over a single discrete channel.

Multichannel audio Audio rendering techniques that utilise two or more discrete audio channels. This includes both stereo and surround-sound systems.

Non-player character (NPC) These are characters within the game who are not directly controlled by a real player and are instead controlled by the game itself. They might be used to progress the games narrative through dialogue, or as digital opponents against which the player must compete. An example would be the ghost enemies in *Pac-Man* [261].

Player avatar The player's virtual representation within the game world. The avatar's actions are directly controlled by the player, allowing the player to navigate through the game world and interact with the objects and systems within it.

Stereo A method for presenting audio over two discrete channels, usually played back over a left and right loudspeaker, or a pair of stereo headphones.

Sound source A game object placed at some position in the game world, from which sound is emitted. This can be diagetic or non-diagetic audio and synthesised or recorded.

Surround-sound A method for presenting audio over more than two discrete audio channels. Examples include 5.1 and 7.1 surround-sound.

Virtual Home Theatre (VHT) A technique for virtualising the loudspeakers of a surround-sound configuration and playing them back over a pair of stereo headphones.

Bibliography

[1] K. Collins, *Game sound: an introduction to the history, theory, and practice of video game music and sound design*. Cambridge, MA: MIT Press, 2008.

[2] M. Kerins, *Beyond Dolby (stereo): cinema in the digital sound age*. Indiana University Press, 2010.

[3] ——, "Multichannel gaming and the aesthetics of interactive surround", in *The Oxford Handbook of New Audiovisual Aesthetics*, J. Richardson, C. Gorbman, and C. Vernallis, Eds. Oxford, UK: Oxford University Press, 2013, pp. 585–605.

[4] D. Williams and N. Lee, *Emotion in Video Game Soundtracking*. Cham, Switzerland: Springer International Publishing, 2018.

[5] Taito, *Space invaders*, Video game, Taito Corporation, 1978.

[6] Department for Digital, Culture, Media and Sport, *Creative industries economic estimates*, `https://www.gov.uk/government/statistics/creative-industries-economic-estimates-january-2016`, [Online; accessed 18-July-2018], 2016.

[7] The Association of UK Interactive Entertainment, *The games industry in numbers*, `https://ukie.org.uk/research`, [Online; accessed 18-July-2018], 2018.

[8] Dolby, *Dolby Atmos*, Surround sound format, Dolby Laboratories, Inc., 2012.

[9] DICE, *Star Wars™ Battlefront™*, Video Game, EA Digital Illusions CE AB, 2015.

[10] Blizzard, *Overwatch*, Video Game, Blizzard Entertainment, Inc., 2016.

[11] J. Schedeen, *Surround sound headphones explained*, `http://uk.ign.com/articles/2010/08/02/surround-sound-headphones-explained`, [Online; accessed 11-April-2018], 2010.

[12] S. Stewart, *Is virtual surround sound worth it for gaming?*, `https://www.gamingscan.com/virtual-surround-sound-worth-it/`, [Online; accessed 11-April-2018], 2018.

[13] S. White, *8 best gaming headsets*, http://www.independent.co.uk/extras/indybest/gadgets-tech/video-games-consoles/best-gaming-headset-pc-ps4-under-100-for-xbox-one-50-a8119911.html, [Online; accessed 11-April-2018], 2017.

[14] M. Andronico and M. Honorof, *Best gaming headsets 2018*, https://www.tomsguide.com/us/best-gaming-headsets,review-2710.html, [Online; accessed 11-April-2018], 2018.

[15] A. Rowe, *Headphone showdown: Virtual surround PC gaming headsets compared!*, https://medium.com/@Xander51/headphone-showdown-virtual-surround-pc-gaming-headsets-compared-526a18d272f, [Online; accessed 11-April-2018], 2017.

[16] B. Hesse, *Best gaming headsets*, https://www.digitaltrends.com/gaming/best-gaming-headsets/, [Online; accessed 11-April-2018], 2018.

[17] S. N. Goodwin, "How players listen", in *Proc of Audio Engineering Society Conference: 35th International Conference: Audio for Games*, London, UK, 2009.

[18] T. Letowski, "Sound quality assessment: Concepts and criteria", in *Proc. of the Audio Engineering Society Convention 87*, Audio Engineering Society, New York, NY, 1989.

[19] F. Rumsey and J. Berg, "Verification and correlation of attributes used for describing the spatial quality of reproduced sound", in *Audio Engineering Society Conference: 19th International Conference: Surround Sound-Techniques, Technology, and Perception*, Audio Engineering Society, Bavaria, Germany, 2001.

[20] J. Berg and F. Rumsey, "Spatial attribute identification and scaling by repertory grid technique and other methods", in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, Artikum, Finland, 1999.

[21] M. Dewhirst, P. Jackson, F. Rumsey, and S. K. Zielinski, "Objective assessement of spatial localisation attributes of surroung-sound reproduction systems", in *Audio Engineering Society Convention 118*, Audio Engineering Society, Barcelona, Spain, 2005.

[22] D. M. Howard and J. Angus, *Acoustics and psychoacoustics (5th Ed.)* New York, NY: Taylor and Francis, 2017.

[23] W. A. Yost, *Fundamentals of hearing: an introduction (5th ed.)* Leiden, The Netherlands: Brill, 2013.

[24] J. O. Pickles, *An introduction to the physiology of hearing (2nd ed.)* London, UK: Academic press Ltd., 1988.

[25] J. Ashmore, *The mechanics of hearing*, `http://www.open.edu/openlearn/ocw/pluginfile.php/654899/mod_resource/content/1/sd329_1_reader1.pdf`, [Online; accessed 27-July-2018], 1989.

[26] A. de Cheveigné, "Pitch perception", in, C. J. Plack, Ed. Oxford, UK: Oxford University Press, 2010, pp. 71–104.

[27] S. W. Rienstra and A. Hirschberg, *An introduction to acoustics*, `http://bbs.hwrf.com.cn/downebd/58308d1309844535-boek.pdf`, [Online; accessed 27-July-2018].

[28] E. Sengpiel, *Sound intensity and the inverse square law*, `http://www.sengpielaudio.com/calculator-squarelaw.htm`, [Online; accessed 1-June-2018].

[29] East Kentucky University, *Decibel (loudness) comparison chart*, `https://music.eku.edu/sites/music.eku.edu/files/ekuhealthandsafety.pdf`, [Online; accessed 17-January-2019].

[30] C. H. Hansen, "Fundamentals of acoustics", *Occupational Exposure to Noise: Evaluation, Prevention and Control. World Health Organization*, pp. 23–52, 2001.

[31] "Timbre", Acoustical Society of America, Definition, 2016.

[32] S. W. Smith, *The scientist and engineer's guide to digital signal processing*. San Diego,CA: California Technical Pub., 1997.

[33] J. O. Smith, *Mathematics of the discrete Fourier transform (DFT): with audio applications*. Stanford, CA: W3K Publishing, 2007.

[34] T. Henderson, *Reflection, refraction, and diffraction*, `http://www.physicsclassroom.com/class/sound/Lesson-3/Reflection,-Refraction,-and-Diffraction`, [Online; accessed 4-June-2018], 2007.

[35] T. Hidakaa, Y. Yamada, and T. Nakagawa, "A new definition of boundary point between early reflections and late reverberation in room impulse responses", *The Journal of the Acoustical Society of America*, vol. 122, no. 1, 2007.

[36] Engineering Toolbox, *Room sound absorption - sound absorption coefficient*, https://www.engineeringtoolbox.com/accoustic-sound-absorption-d_68.html, [Online; accessed 7-June-2018], 2003.

[37] Acoustic Traffic, *Absorption coefficients*, http://www.acoustic.ua/st/web_absorption_data_eng.pdf, [Online; accessed 17-January-2019].

[38] S. A. Gelfand, *Hearing: An introduction to psychological and physiological acoustics (6th ed.)* Boca Raton, FL: CRC Press, 2016.

[39] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect", *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999.

[40] D. M. Howard and D. T. Murphy, *Voice science, acoustics, and recording*. San Diego, CA: Plural Publishing, 2007.

[41] B. C. Moore, *An introduction to the psychology of hearing (6th ed.)* Leiden, The Netherlands: Brill, 2013.

[42] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity", *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2110–2117, 2002.

[43] T. Neher and T. Brookes, "Training of listeners for the evaluation of spatial sound reproduction", in *AES Convention Papers*, AES, 2002, p. 5584.

[44] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique", in *Audio Engineering Society Convention 108*, Audio Engineering Society, Paris, France, 2000.

[45] ——, "Advancements in impulse response measurements by sine sweeps", in *Audio Engineering Society Convention 122*, Audio Engineering Society, Vienna, Austria, 2007.

[46] F. Rumsey, *Spatial Audio*. Oxford, UK: Focal Press, 2001.

[47] R. S. Woodworth and H. Schlosberg, *Experimental Psychology. New York: Holt, 1938*. Oxford, UK: Oxford and IBH Publishing.

[48] A. A. Scharine and T. R. Letowski, *Factors affecting auditory localization and situational awareness in the urban battlefield*, http://www.dtic.mil/dtic/tr/fulltext/u2/a431963.pdf, [Online; accessed 27-July-2018], 2005.

[49] C. Searle, L. Braida, M. Davis, and H. Colburn, "Model for auditory localization", *The Journal of the Acoustical Society of America*, vol. 60, no. 5, pp. 1164–1175, 1976.

[50] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E Korner, "A probabilistic model for binaural sound localization", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 5, pp. 982–994, 2006.

[51] D. McAlpine, "Creating a sense of auditory space", *The Journal of physiology*, vol. 566, no. 1, pp. 21–28, 2005.

[52] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. Cambridge, MA: MIT Press, 1983.

[53] L. Rayleigh, "On our perception of sound direction", *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907.

[54] Sweetwater, *Comb filter*, `https://www.sweetwater.com/insync/comb-filter/`, [Online; accessed 17-January-2019], 1998.

[55] G. S. Kendall and W. L. Martens, "Simulating the cues of spatial hearing in natural environments", in *Proc. of International Computer Music Conference*, 1984, pp. 111–125.

[56] D. R. Begault and L. J. Trejo, *3-D sound for virtual reality and multimedia*. Academic Press Inc., 2000.

[57] B. Xie, *Head-related transfer function and virtual auditory display*. J. Ross Publishing, 2013.

[58] G. Kearney, C. Armstrong, and L. Thresh, *Sadie ii database*, `https://www.york.ac.uk/sadie-project/database.html`, [Online; accessed 01-July-2018], 2018.

[59] U. Horbach and M. M. Boone, "Future transmission and rendering formats for multichannel sound", in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, Audio Engineering Society, Artikum, Finland, 1999.

[60] G. Kearney and T. Doyle, "An HRTF database for virtual loudspeaker rendering", in *Audio Engineering Society Convention 139*, Audio Engineering Society, New York, NY, 2015.

[61] H. Møller, "Fundamentals of binaural technology", *Applied acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992.

[62] B. Masiero, J. Fels, and M. Vorländer, "Review of the crosstalk cancellation filter technique", in *International Conference on Spatial Audio (ICSA)*, Tonmeister, Detmold, Germany, 2011.

[63] E. Y. Choueiri, "Optimal crosstalk cancellation for binaural audio with two loudspeakers", *Princeton University*, vol. 28, 2008.

[64] D. R. Begault, "Challenges to the successful implementation of 3-d sound", *Journal of the Audio Engineering Society*, vol. 39, no. 11, pp. 864–870, 1991.

[65] W. G. Gardner and K. D. Martin, "HRTF measurements of a kemar", *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.

[66] B. Gardner, K. Martin, *et al.*, *Hrft measurements of a kemar dummy-head microphone*, 1994.

[67] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database", in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, IEEE, 2001, pp. 99–102.

[68] IRCAM, *Listen HRTF database*, `http://recherche.ircam.fr/equipes/salles/listen/`, [Online; accessed 01-July-2018], 2003.

[69] G.R.A.S, *Gras 45bb-1 kemar head and torso for hearing aid test, 1-ch lemo*, `https://www.gras.dk/products/head-torso-simulators-kemar/kemar-for-hearing-aid-test-1-ch/product/499-45bb-1`, [Online; accessed 01-July-2018], 2018.

[70] D. J. Robinson and R. G. Greenfield, "A binaural simulation which renders out-of-head localization with low-cost digital signal processing of head-related-transfer functions and pseudo reverberation", in *Audio Engineering Society Convention 104*, Audio Engineering Society, Amsterdam, The Netherlands, 1998.

[71] R. Crawford-Emery and H. Lee, "The subjective effect of BRIR length on perceived headphone sound externalization and tonal coloration", in *Audio Engineering Society Convention 136*, Audio Engineering Society, Berlin, Germany, 2014.

[72] K. Collins, *Playing with sound: a theory of interacting with sound and music in video games*. Cambridge, MA: MIT Press, 2013.

[73] 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, pp. 651–666, 2002.

[74] B. Gardner, *Maintaining mono compatibility in your music mix*, https://resoundsound.com/mono-compatibility/, [Online; accessed 18-January-2019], 2013.

[75] P. White, *Making the most of the stereo panorama*, https://www.soundonsound.com/techniques/making-most-stereo-panorama, [Online; accessed 18-January-2019], 2009.

[76] S. P. Lipshitz, "Stereo microphone techniques: Are the purists wrong?", *Journal of the Audio Engineering Society*, vol. 34, no. 9, pp. 716–744, 1986.

[77] A. D. Blumlein, "British patent specification 394,325 (improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems)", *Journal of the Audio Engineering Society*, vol. 6, no. 2, pp. 91–130, 1958.

[78] M. A. Gerzon and J. M. Woram, "Blumlein stereo microphone technique-and author's reply", *Journal of the Audio Engineering Society*, vol. 24, no. 1, pp. 36–38, 1976.

[79] F. Rumsey and T. McCormick, *Sound and recording: an introduction (6th ed.)* Oxford, UK: Focal Press, 2009.

[80] B. B. Bauer, "Phasor analysis of some stereophonic phenomena", *The Journal of the Acoustical Society of America*, vol. 33, no. 11, pp. 1536–1539, 1961.

[81] V. Pulkki and M. Karjalainen, "Localization of amplitude-panned virtual sources 1: Stereophonic panning", *Journal of the Audio Engineering Society*, vol. 49, no. 9, pp. 739–752, 2001.

[82] Dolby Laboratories, *Dolby surround mixing manual*, http://www.associationdesmixeurs.fr/wp-content/uploads/2015/10/Dolby-Surround-Mix-Manual.pdf, [Online; accessed 18-January-2019], 1998.

[83] Dolby, *Dolby stereo*, Surround-sound format, Dolby Laboratories, Inc., 1975.

[84] ——, *Dolby surround*, Surround-sound format, Dolby Laboratories, Inc., 1982.

[85] L. Blake, "Mixing Dolby stereo film sound", *Recording Engineer/Producer*, vol. 12, no. 1, 1981.

[86] ITU-R, "775-2, multi-channel stereophonic sound system with or without accompanying picture", *Telecommunications Union Radiocommunication Assembly, Geneva, Switzerland*, vol. 2004, 2004.

[87] T. Holman, *Surround sound: up and running (2nd ed.)* Burlington, MA: Focal Press, 2014.

[88] R. C. Cabot, "Sound localization in 2 and 4 channel systems: A comparison of phantom image prediction equations and experimental data", in *Audio Engineering Society Convention 58*, Audio Engineering Society, New York, NY, 1977.

[89] G. Martin, W. Woszczyk, J. Corey, and R. Quesnel, "Sound source localization in a five-channel surround sound reproduction system", in *Audio Engineering Society Convention 107*, Audio Engineering Society, New York, NY, 1999.

[90] G. Theile and G. Plenge, "Localization of lateral phantom sources", *Journal of the Audio Engineering Society*, vol. 25, no. 4, pp. 196–200, 1977.

[91] G. Martin, W. Woszczyk, J. Corey, and R. Quesnel, "Controlling phantom image focus in a multichannel reproduction system", in *Audio Engineering Society Convention 107*, Audio Engineering Society, New York, NY, 1999.

[92] C. E. Shannon, "Communication in the presence of noise", *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[93] Chegg, *Aliasing*, `https://www.chegg.com/homework-help/definitions/aliasing-4`, [Online; accessed 21-January-2019], 2019.

[94] DTS, *DTS HD master audio*, Surround-sound format, Dolby Laboratories, Inc., 2013.

[95] Dolby, *Dolby true HD*, Surround-sound format, Dolby Laboratories, Inc., 2013.

[96] Dolby Laboratories, Inc, "Dolby metadata guide", Tech. Rep. Issue 3, 2005.

[97] H. Robjohns, *Surround sound explained: Part 5*, `https://www.soundonsound.com/techniques/surround-sound-explained-part-5`, [Online; accessed 11-July-2018], 2001.

[98] A. Digenis, "Challenges of the headphone mix in games", in *Audio Engineering Society Conference: 56th International Conference: Audio for Games*, Audio Engineering Society, London, UK, 2015.

[99] N. Zacharov and J. Huopaniemi, "Results of a round robin subjective evaluation of virtual home theatre sound systems", in *Audio Engineering Society Convention 107*, Audio Engineering Society, New York, NY, 1999.

[100] N. Zacharov and G. Lorho, "Subjective evaluation of virtual home theatre sound systems for loudspeakers and headphones", in *Audio Engineering Society Convention 116*, Audio Engineering Society, Berlin, Germany, 2004.

[101] R. Mason and F. Rumsey, "An assessment of spatial performance of virtual home theatre algorithms by subjective and objective methods", *Audio Engineering Society Preprint*, vol. 5137, 2000.

[102] Turtle Beach, *Ear force i60*, `https://shop.turtlebeach.com/us/i60`, [Online; accessed 4-September-2017], 2017.

[103] Razer, *Razer surround*, `https://www.razerzone.com/gb-en/surround`, [Online; accessed 4-September-2017], 2017.

[104] M. A. Gerzon, "Periphony: With-height sound reproduction", *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[105] Brilliant.org, *Spherical harmonics*, `https://brilliant.org/wiki/spherical-harmonics/`, [Online; accessed 21-January-2019], 2019.

[106] D. Malham, "Higher order ambisonic systems", *abstracted from 'Space in Music-Music in Space', an MPhil thesis by Dave Malham, submitted to the University of York in April*, 2003.

[107] M. Kronlachner, "Plug-in suite for mastering the production and playback in surround sound and ambisonics", *Gold-Awarded Contribution to AES Student Design Competition*, 2014.

[108] J.-M. Jot and O. Warusfel, "Spat: A spatial processor for musicians and sound engineers", in *CIARM: International Conference on Acoustics and Musical Research*, 1995.

[109] *Ambisonic channels*, accessed 09/12/2016. [Online]. Available: `http://ambisonics.ch/standards/channels/`.

[110] D. G. Malham and A. Myatt, "3-D sound spatialization using ambisonic techniques", *Computer Music Journal*, vol. 19, no. 4, pp. 58–70, 1995.

[111] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "Ambix-a suggested ambisonics format", in *in Proc. of Ambisonics Symposium, Lexington*, 2011.

[112] 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, Audio Engineering Society, Saariselka, Finland, 2007.

[113] M. Frank, F. Zotter, H. Wierstorf, and S. Spors, "Spatial audio rendering", in *Quality of Experience*, Springer, 2014, pp. 247–260.

[114] E. Bates, G. Kearney, D. Furlong, and F. Boland, "Localization accuracy of advanced spatialisation techniques in small concert halls", *Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 3069–3069, 2007.

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

[116] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory (SAQI)", *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 984–994, 2014.

[117] S. Le Bagousse, M. Paquier, and C. Colomes, "Families of sound attributes for assessment of spatial audio", in *129th AES Convention*, San Francisco, CA, 2010.

[118] S. Le Bagousse, M. Paquier, C. Colomes, and S. Moulin, "Sound quality evaluation based on attributes-application to binaural contents", in *Audio Engineering Society Convention 131*, Audio Engineering Society, New York, NY, 2011.

[119] S. Bech and N. Zacharov, *Perceptual audio evaluation-Theory, method and application*. Chichester, UK: John Wiley and Sons, 2007.

[120] P. Zahorik, "Auditory display of sound source distance", in *in Proc. of International Conference on Auditory Display*, 2002, pp. 326–332.

[121] F. Rumsey, B. de Bruyn, and N. Ford, "Graphical elicitation techniques for subjective assessment of the spatial attributes of loudspeaker reproduction a pilot investigation", in *Audio Engineering Society Convention 110*, Audio Engineering Society, Amsterdam, The Netherlands, 2001.

[122] R. Nicol, L. Gros, C. Colomes, M. Noisternig, O. Warusfel, H. Bahu, B. F. Katz, and L. S. Simon, "A roadmap for assessing the quality of experience of 3d audio binaural rendering", *10.14279/depositonce-4103*, 2014.

[123] T. Okano, L. L. Beranek, and T. Hidaka, "Relations among interaural cross-correlation coefficient (IACC), lateral fraction (LF), and apparent source width (ASW) in concert halls", *The Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 255–265, 1998.

[124] W. J. Cavanaugh, G. C. Tocci, and J. A. Wilkes, *Architectural acoustics: Principles and practice*. New York, NY: John Wiley and Sons, 2010.

[125] K. B. Christensen and T. Lund, "Room simulation for multichannel film and music", in *Audio Engineering Society Convention 107*, Audio Engineering Society, New York, NY, 1999.

[126] S. Choisel and F. Wickelmaier, "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference", *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 388–400, 2007.

[127] Giant Bomb, *Space invaders*, `https://www.giantbomb.com/space-invaders/3030-5099/`, [Online; accessed 21-January-2019], 2019.

[128] Magnavox, *Odyssey*, Video game console, Magnavox, 1972.

[129] Atari, *Pong*, Video game console, Atari Inc., 1972.

[130] C. Shea, *Al alcorn interview*, `http://uk.ign.com/articles/2008/03/11/al-alcorn-interview`, [Online; accessed 03-February-2018], 2008.

[131] Commodore, *Amiga 1000*, Video game console, Commodore Int., Ltd., 1985.

[132] J. Weske, *Digital sound and music in computer games*, `http://3daudio.info/gamesound/history.html`, [Online; accessed 19-September-2017], 2015.

[133] Wikipedia, *Screenshot of Pong*, `https://en.wikipedia.org/wiki/File:Pong.png`, [Online; accessed 19-July-2018], 2006.

[134] Nintendo, *Super Nintendo entertainment system (SNES)*, Video game console, Nintendo Co., Ltd., 1991.

[135] C. Lara, *Throwback Thursday: Top 5 uses of surround sound on SNES*, `https://www.gamnesia.com/articles/throwback-thursday-top-5-uses-of-surround-sound-on-snes`, [Online; accessed 19-September-2017], 2013.

[136] H. Robjohns, "Surround sound explained: Part 2", *Sound on Sound 16 (11)*, pp. 170 –176, 2001, [Magazine].

[137] M. P. Simpson, *The first King Arthur, and Dolby Surround*, `http://www.martinpsimpson.com/2011/09/first-king-arthur-and-dolby-surround.html`, [Online; accessed 16-September-2017], 2011.

[138] Ocean, *Jurassic Park*, Video game, Ocean Software Ltd., 1993.

[139] Argonaut Games, *Vortex*, Video game, Argonaut Games PLC, 1994.

[140] SNK, *Samurai Shodown*, Video game, SNK Corp., 1993.

[141] Argonaut Games, *King Arthur's world*, Video game, Argonaut Games PLC, 1993.

[142] Game Fabrique, *Screenshot of King Arthur's World*, https://tinyurl.com/y8rw4fzp, [Online; accessed 19-July-2018].

[143] Sony, *Playstation 2 (PS2)*, Video game console, Sony Computer Entertainment, 2000.

[144] P. De Lancie, *Sound for Playstation 2*, https://www.mixonline.com/sfp/sound-playstation2-369087, [Online; accessed 9-May-2018], 2001.

[145] DTS, *DTS Interactive*, Surround-sound format, DTS, Inc., 2001.

[146] Rockstar North, *Grand theft auto: Vice city*, Video game, Rockstar North Limited, 2002.

[147] EA Canada, *FIFA soccer 2003*, Video game, Electronic Arts Inc. (EA), 2002.

[148] Moby Games, *Sound capabilities: Dts*, http://www.mobygames.com/attribute/sheet/attributeId,207/p,7/, [Online; accessed 9-May-2018], 2018.

[149] Dolby, *Dolby digital live*, https://www.dolby.com/us/en/technologies/dolby-digital-live.html, [Online; accessed 9-May-2018], 2018.

[150] Bungie, *Halo: Combat evolved*, Video game, Bungie, Inc., 2001.

[151] Big Blue Box Studios, *Fable*, Video game, Lionhead Studios Ltd., 2004.

[152] Bethesda Game Studios, *The Elder Scrolls III: Morrowind*, Video game, Bethesda Softworks LLC, 2002.

[153] Sony, *Playstation 3 (PS3)*, Video game console, Sony Computer Entertainment, 2006.

[154] ——, *Playstation 4 (PS4)*, Video game console, Sony Interactive Entertainment, 2013.

[155] Microsoft, *Xbox one*, Video game console, Microsoft Corp., 2013.

[156] R. Waniata, *PS4 is first to support new DTS-HD decoder, proving the console is about more than just looks*, https://www.digitaltrends.com/home-theater/ps4s-inclusion-dts-hd-decoding-proves-just-looks/, [Online; accessed 25-May-2018], 2013.

[157] Medium, *Screenshot of halo: Combat evolved*, https://medium.com/@Oozer3993/the-making-of-halo-how-combat-evolved-from-blam-part-1-f6b58fcc4ade, [Online; accessed 19-July-2018], 2016.

[158] Sound Blaster, *About sound blaster*, http://www.soundblaster.com/about/, [Online; accessed 25-May-2018], 2016.

[159] Mule Soft, *What is an API? (application programming interface)*, https://www.mulesoft.com/resources/api/what-is-an-api, [Online; accessed 25-May-2018], 2018.

[160] Microsoft, *Direct X*, Audio API, Microsost Corporation, 1995.

[161] Creative Technology, *Openal*, Audio API, Creative Technology Ltd., 2009.

[162] ——, *A3d*, Audio API, Creative Technology Ltd., 1997.

[163] Blue Ripple Sound, *Welcome to blue ripple sound*, http://www.blueripplesound.com/, [Online; accessed 25-May-2018], 2015.

[164] 2k Games, *Bioshock*, Video game, 2K Games, Inc., 2007.

[165] id Software, *Doom 3*, Video game, id Software LLC, 2004.

[166] Rockstar Games, *Grand theft auto v*, Video game, Rockstar Games, Inc., 2013.

[167] Team Ico, *The last guardian*, Video game, Sony Interactive Entertainment, LLC., 2016.

[168] Somethin' Else, *Papa sangre 2*, Video Game, Somethin' Else, 2013.

[169] Ninja Theory, *Hellblade: Seunua's sacrifice*, Video Game, Ninja Theory, Ltd., 2017.

[170] BAFTA, *Audio achievement in 2018*, http://awards.bafta.org/award/2018/games/audio-achievement, [Online; accessed 24-May-2018], 2018.

[171] J. Ramos, *Hellblade is an audio nightmare*, https://www.polygon.com/2017/8/9/16120082/hellblade-binaural-audio-psychosis, [Online; accessed 05-February-2017], 2017.

[172] Codemasters, *Colin McRae: Dirt*, Video Game, Codemasters Software Ltd., 2007.

[173] Audiokinetic, *Ambisonics in Wwise: Overview*, https://www.audiokinetic.com/products/ambisonics-in-wwise/, [Online; accessed 20-September-2017], 2016.

[174] Google, *Discover resonance audio*, https://developers.google.com/resonance-audio/discover/overview, [Online; accessed 23-May-2018], 2017.

[175] Infinity Ward, *Call of duty 4: Modern warfare*, Video game, Infinity Ward, Inc., 2007.

[176] Valve Corporation, *Portal 2*, Video game, Valve Corporation, 2011.

[177] Bossa Studios, *Surgeon simulator*, Video game, Bossa Studios Ltd., 2013.

[178] Mojang, *Minecraft*, Video game, Microsoft Studios, 2011.

[179] The Video Game Gallery, *Screenshot of doom (2016)*, `http://www.thevideogamegallery.com/gallery/image:19913/doom-2016:unwilling-cacodemon`, [Online; accessed 19-July-2018], 2015.

[180] id Software, *Doom*, Video game, id Software LLC, 2016.

[181] Nintendo, *Super mario 64*, Video game, Nintendo Co., Ltd., 1996.

[182] Core Design, *Tomb raider*, Video game, Core Design, Ltd., 1996.

[183] Epic Games, *Gears of war*, Video game, Epic Games, Inc., 2006.

[184] thatgamecompany, *Journey*, Video game, thatgamecompany, LLC., 2015.

[185] Kotaku, *Screenshot of the witcher 3: Wild hunt*, `https://kotaku.com/the-witcher-3-as-told-through-beautiful-screenshots-1723020971`, [Online; accessed 19-July-2018], 2015.

[186] Monolith Productions, *Middle Earth: Shadow of Mordor*, Video game, Warner Bros. Interactive Entertainment, 2014.

[187] Insomniac Games, *Ratchet and Clank*, Video game, Insomniac Games, Inc., 2016.

[188] Platinum Games, *Nier: Automata*, Video game, Platinum Games Inc., 2017.

[189] Nintendo, *Wii*, Video game console, Nintendo Co., Ltd., 2006.

[190] Crystal Dynamics, *Rise of the tomb raider*, Video game, Square Enix Holdings Co., Ltd., 2015.

[191] Bluepoint Games and Team Ico, *The ico and shadow of the colossus collection*, Video game, Bluepoint Games, Inc., 2011.

[192] thatgamecompany, *Flower*, Video game, thatgamecompany, LLC., 2009.

[193] Playtonic Games, *Yooka laylee*, Video game, Playtonic Limited, 2017.

[194] Vicarious Visions, *Crash bandicoot n sane trilogy*, Video game, Vicarious Visions, Inc., 2017.

[195] Koei Tecmo, *Attack on titan: Wings of freedom*, Video game, Koei Tecmo Holdings Co., Ltd., 2016.

[196] The Chinese Room, *Dear esther: Landmark edition*, Video game, The Chinese Room Ltd., 2017.

[197] Guerilla Games, *Horizon zero dawn*, Video game, Sony Interactive Entertainment, LLC, 2017.

[198] Naught Dog, *The last of us: Remastered*, Video Game, Naughty Dog, LLC, 2014.

[199] DualShockers, *Naughty dog's the last of us is the most awarded game in recorded history by critics*, http://www.dualshockers.com/the-last-of-us-is-the-most-awarded-game-in-history-by-critics/, [Online; accessed 21-August-2017], 2014.

[200] BAFTA, *Audio achievement in 2014*, http://awards.bafta.org/award/2014/games/audio-achievement, [Online; accessed 24-May-2018], 2014.

[201] Game Audio Network Guild, *Gang awards 2013 final winners*, http://www.audiogang.org/awards/2014-awards/, [Online; accessed 24-May-2018], 2014.

[202] Creative Assembly, *Alien: Isolation*, Video game, The Creative Assembly Limited, 2014.

[203] Internet Movie Database, *Awards (alien 1979)*, https://www.imdb.com/title/tt0078748/awards, [Online; accessed 25-May-2018], 2018.

[204] J. Macgregor, *Seeing with your ears — the audio of alien: Isolation*, https://www.pcgamer.com/the-audio-of-alien-isolation/, [Online; accessed 25-May-2018], 2015.

[205] R. Roberts, "Fear of the unknown: Music and sound design in psychological horror games", in *Music In Video Games*, Routledge, 2014, pp. 152–164.

[206] Kojima Productions, *P.T.* Video game, Kojima Productions Co., Ltd. (no longer available as a download), 2014.

[207] M. Kamen, *Kojima: Next silent hill will 'make you sh*t your pants'*, http://www.wired.co.uk/article/kojima-talks-on-pt-and-silent-hills, [Online; accessed 26-May-2018], 2014.

[208] Playstation, *Screenshot of ratchet and clank: Tools of destruction*, http://www.thevideogamegallery.com/gallery/image:19913/doom-2016:unwilling-cacodemon, [Online; accessed 19-July-2018].

[209] U. Jekosch, *Voice and speech quality perception: assessment and evaluation*. Berlin, Germany: Springer-Verlag Berlin Heidelberg, 2006.

[210] A. Raake and S. Egger, "Quality and quality of experience", in *Quality of experience*, Cham, Switzerland: Springer, 2014, pp. 11–33.

[211] D. Kahneman, "Objective happiness", *Well-being: The foundations of hedonic psychology*, vol. 3, p. 25, 1999.

[212] P. Lebreton, A. Raake, M. Barkowsky, and P. Le Callet, "Perceptual preference of S3D over 2D for HDTV in dependence of video quality and depth", in *IVMSP Workshop, 2013 IEEE 11th*, IEEE, Seoul, South Korea, 2013, pp. 1–4.

[213] A. Raake and C. Schlegel, "Auditory assessment of conversational speech quality of traditional and spatialized teleconferences", in *Voice Communication (Sprachkommunikation), 2008 ITG Conference on*, VDE, Aachen, Germany, 2008, pp. 1–4.

[214] A. Wilson and B. Fazenda, "Relationship between hedonic preference and audio quality in tests of music production quality", in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*, IEEE, Lisbon, Portugal, 2016, pp. 1–6.

[215] S. Choisel and F. Wickelmaier, "Relating auditory attributes of multichannel reproduced sound to preference and to physical parameters", in *Audio Engineering Society Convention 120*, Audio Engineering Society, Paris, France, 2006.

[216] H. A. David, *The method of paired comparisons (2nd ed.)* London, UK: Oxford University Press, 1988.

[217] H. Scheffé, "An analysis of variance for paired comparisons", *Journal of the American Statistical Association*, vol. 47, no. 259, pp. 381–400, 1952.

[218] I. Recommendation, "General methods for the subjective assessment of sound quality", *ITU-R BS*, pp. 1284–1, 2003.

[219] I. T. Union, *Recommendation ITU-R BS. 1534-3: Method for the subjective assessment of intermediate quality level of audio systems*, 2014.

[220] G. A. Soulodre and M. C. Lavoie, "Subjective evaluation of large and small impairments in audio codecs", in *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*, Audio Engineering Society, Florence, Italy, 1999.

[221] "Qoe".

[222] K. Brunnström, S. A. Beker, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, M.-C. Larabi, *et al.*, "Qualinet white paper on definitions of quality of experience", 2013.

[223] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, "Factors influencing quality of experience", in *Quality of experience*, Springer, 2014, pp. 55–72.

[224] A. Raake, H. Wierstorf, and J. Blauert, "A case for TWO! ears in audio quality assessment", in *Forum Acusticum*, 2014, p. 41.

[225] J. G. Beerends and F. E. De Caluwe, "The influence of video quality on perceived audio quality and vice versa", *Journal of the Audio Engineering Society*, vol. 47, no. 5, pp. 355–362, 1999.

[226] E. T. Davis, K. Scott, J. Pair, L. F. Hodges, and J. Oliverio, "Can audio enhance visual perception and performance in a virtual environment?", in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications Sage CA: Los Angeles, CA, vol. 43, 1999, pp. 1197–1201.

[227] F. N. Rahayu, U. Reiter, J. You, A. Perkis, and T. Ebrahimi, "Subjective visual quality assessment in the presence of audio for digital cinema", in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, IEEE, Mechelen, Belgium, 2011, pp. 113–118.

[228] R. B. Welch and D. H. Warren, "Immediate perceptual response to intersensory discrepancy.", *Psychological Bulletin*, vol. 88, no. 3, p. 638, 1980.

[229] M. P. Hollier and R Voelcker, "Objective performance assessment: Video quality as an influence on audio perception", in *Audio Engineering Society Convention 103*, Audio Engineering Society, New York, NY, 1997.

[230] J. Lassalle, L. Gros, and G. Coppin, "Combination of physiological and subjective measures to assess quality of experience for audiovisual technologies", in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, IEEE, Mechelen, Belgium, 2011, pp. 13–18.

[231] N. F. Dixon and L. Spitz, "The detection of auditory visual desynchrony", *Perception*, vol. 9, no. 6, pp. 719–721, 1980.

[232] S. Komiyama, "Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems", *Journal of the Audio Engineering Society*, vol. 37, no. 4, pp. 210–214, 1989.

[233] T. De Pessemier, K. De Moor, A. J. Verdejo, D. Van Deursen, W. Joseph, L. De Marez, L. Martens, and R. Van de Walle, "Exploring the acceptability of the audiovisual quality for a mobile video session based on objectively measured

parameters", in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, Mechelen, Belgium, 2011, pp. 125–130.

[234] R. K. Mok, E. W. Chan, and R. K. Chang, "Measuring the quality of experience of http video streaming", in *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*, IEEE, Dublin, Ireland, 2011, pp. 485–492.

[235] M. R. Quintero and A. Raake, "Towards assigning value to multimedia QoE", in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, IEEE, Mechelen, Belgium, 2011, pp. 1–6.

[236] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality", *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 968–976, 2005.

[237] S. K. Zielinski, F. Rumsey, R. Kassier, and S. Bech, "Comparison of basic audio quality and timbral and spatial fidelity changes caused by limitation of bandwidth and by down-mix algorithms in 5.1 surround audio systems", *Journal of the Audio Engineering Society*, vol. 53, no. 3, pp. 174–192, 2005.

[238] C. Pike and F. Melchior, "An assessment of virtual surround sound systems for headphone listening of 5.1 multichannel audio", in *Audio Engineering Society Convention 134*, Audio Engineering Society, Rome, Italy, 2013.

[239] F. W. J. Sousa, "Subjective comparison between stereo and binaural processing from b-format ambisonic raw audio material", in *Audio Engineering Society Convention 130*, Audio Engineering Society, London, UK, 2011.

[240] R. van Tol and S. Huiberts, *Ieza: A framework for game audio*, `https://www.gamasutra.com/view/feature/131915/ieza_a_framework_for_game_audio.php`, [Online; accessed 27-July-2018], 2008.

[241] S. K. Zielinski, F. Rumsey, S. Bech, B. De Bruyn, and R. Kassier, "Computer games and multichannel audio quality-the effect of division of attention between auditory and visual modalities", in *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*, Audio Engineering Society, Banff, Canada, 2003.

[242] A. Field and G. Hole, *How to design and report experiments*. London, UK: Sage Publications Ltd., 2003.

[243] S. Siegal and N. J. Castellan, *Nonparametric statistics for the behavioral sciences (2nd ed.)* New York, NY: McGraw-hill, 1988.

[244] S. Werner, J. Liebetrau, and T. Sporer, "Audio-visual discrepancy and the influence on vertical sound source localization", in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, IEEE, Yarra Valley, Australia, 2012, pp. 133–139.

[245] Cycling '74, *Max/msp*, Computer Software, Cycling '74, 2017.

[246] E. Ltd, *Edtracker pro wired*, `https://www.edtracker.co.uk/shop/my-basket/edtracker-pro`, [Online; accessed 18-April-2018], 2018.

[247] J. Berg and F. Rumsey, "Identification of quality attributes of spatial audio by repertory grid technique", *Journal of the Audio Engineering Society*, vol. 54, no. 5, pp. 365–379, 2006.

[248] Unity, *Unity game engine*, Video Game Engine, Unity Technologies SF, 2005.

[249] J Postel, *User datagram protocol*, `https://tools.ietf.org/html/rfc768`, [Online; accessed 18-July-2018], 1980.

[250] Y. Suzuki, D Brungart, H Kato, K Iida, D Cabrera, and Y Iwaya, *Principles and Applications of Spatial Hearing*. World Scientific, 2011.

[251] R. N. Buliung, K. Larsen, G. E. Faulkner, and M. R. Stone, "The "path" not taken: Exploring structural differences in mapped-versus shortest-network-path school travel routes", *American journal of public health*, vol. 103, no. 9, pp. 1589–1596, 2013.

[252] 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, no. 6, pp. 451–469, 1996.

[253] A. W. Bronkhorst, "Localization of real and virtual sound sources", *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2542–2553, 1995.

[254] Unity, *Audio listener*, `https://docs.unity3d.com/Manual/class-AudioListener.html`, [Online; accessed 9-May-2018].

[255] Audiokinetic, *Wwise*, Video game audio middleware, Audiokinetic Inc., 2006.

[256] Firelight Technologies, *Fmod studio*, Video game audio middleware, Firelight Technologies Pty, Ltd., 2017.

[257] Epic Games, *Unreal engine 4*, Video game engine, Epic Games, Inc., 2014.

[258] Square Enix, *Luminous engine*, Video game engine, Square Enix Holdings Co., Ltd., 2012.

[259] CD Projekt Red, *Redengine*, Video game engine, CD Projekt S.A., 2011.

[260] Unity, *Gameobjects*, `https://docs.unity3d.com/Manual/GameObjects.html`, [Online; accessed 28-March-2017].

[261] Namco, *Pac-man*, Video game, Namco Limited, 1980.