THE QUALITY OF EXPERIENCE OF NEXT GENERATION AUDIO: EXPLORING SYSTEM, CONTEXT AND HUMAN INFLUENCE FACTORS

TIM WALTON

In partial fulfilment of the requirements
for the degree of Doctor of Philosophy

School of Computing
Newcastle University

BBC Research & Development

May 2018

ABSTRACT

The next generation of audio reproduction technology has the potential to deliver immersive and personalised experiences to the user; multichannel with-height loudspeaker arrays and binaural techniques offer 3D audio experiences, whereas object-based techniques offer possibilities of adapting content to suit the system, context and user. A fundamental process in the advancement of such technology is perceptual evaluation. It is crucial to understand how listeners perceive new technology in order to drive future developments. This thesis explores the experience provided by next generation audio technology by taking a quality of experience (QoE) approach to evaluation. System, context and human factors all influence QoE and in this thesis three case studies are presented to explore the role of these categories of influence factors (IFs) in the context of next generation audio evaluation. Furthermore, these case studies explore suitable methods and approaches for the evaluation of the QoE of next generation audio with respect to its various IFs. Specific contributions delivered from these individual studies include a subjective comparison between soundbar and discrete surround sound technology, the application of the Open Profiling of Quality method to the field of audio evaluation, an understanding of both how and why environmental noise influences preferred audio object balance, an understanding of how the influence of technical audio quality on overall listening experience is related to a range of psychographic variables and an assessment of the impact of binaural processing on overall listening experience. When considering these studies as a whole, the research presented here contributes the thesis that to effectively evaluate the perceived quality of next generation audio, a QoE mindset should be taken that considers system, context and human IFs.

Dedicated to the memory of my father, Robert Walton.

ACKNOWLEDGEMENTS

The research presented in this thesis was only possible thanks to the support from the following people. Firstly, I would like to express gratitude to the Engineering and Physical Sciences Research Council (EPSRC) and the British Broadcasting Corporation Research & Development department (BBC R&D), whose funding made this project possible [grant number EP/L505560/1]. For your support, expertise and insight I am deeply grateful to Michael Evans, David Kirk, Frank Melchior and Peter Wright; your guidance and feedback throughout this project has been invaluable. Thank you to colleagues at BBC R&D for your knowledge and technical assistance with the practical aspects of this research; it is inspiring to work alongside those who are at the forefront of the field with regards to both knowledge and enthusiasm. I am extremely grateful to all of those who participated in the studies; this research really would not have been possible without you. Thanks also go to the examiners of this thesis, Francis Rumsey and Kyle Montague, for your thoughtful and constructive feedback. Last but not least, thank you to my family, to my friends and to Linda; you have inspired and encouraged me to embark on this journey and supported me throughout.

Tim Walton

Bewdley, March 2018

CONTENTS

ABSTRACT iii

ACKNOWLEDGEMENTS vii

LIST OF FIGURES xiv

LIST OF TABLES xvii

1 INTRODUCTION 1

1.1 Background and Motivation . . . 1

1.2 Aims and Objectives . . . 4

1.3 Original Contributions . . . 6

1.3.1 List of Publications . . . 7

1.3.2 Data Access . . . 8

1.4 Structure of the Thesis . . . 9

2 AUDIO TECHNOLOGY AND ITS EVALUATION 11

2.1 Introduction . . . 11

2.2 From Mono to Next Generation . . . 11

2.2.1 Reproduction in the Horizontal Plane . . . 12

2.2.2 Immersive Audio . . . 13

2.2.3 Object-Based Audio . . . 14

2.2.4 Next Generation Audio: A Summary . . . 17

2.3 Considerations on the Term "Quality" . . . 18

2.3.1 Definitions of "Quality" . . . 19

2.3.2 The Quality-Formation Process . . . 20

2.3.3 Sound Quality . . . 20

2.4 Audio Quality Evaluation Principles . . . 23

2.4.1 Physical, Perceptual and Affective Measurement . . . 23

2.4.2 Validity . . . 25

2.4.3 Assessor Categorisation . . . 26

2.5 Global Judgment Methods . . . 28

2.5.1 ITU-R BS.1116 . . . 29

2.5.2 ITU-R BS.1534 (MUSHRA) . . . 30

2.5.3 ITU-T P.800 . . . 31

2.5.4 Affective Measures . . . 32
2.6 Attribute-Based Methods . . . 33
2.6.1 Considerations on Attribute Selection . . . 36
2.6.2 Provided Construct Methods . . . 37
2.6.3 Elicited Construct Methods . . . 38
2.6.4 An Overview of Perceptual Attributes . . . 41
2.7 Combining Global and Attribute-Based Methods . . . 44
2.8 Sound Quality Models . . . 45
2.9 Evaluation of Next Generation Audio: The State of the Art . . . 46
2.9.1 Immersive Audio . . . 46
2.9.2 Object-Based Audio . . . 48
2.10 Discussion . . . 49
2.11 Summary . . . 50
3 QUALITY OF EXPERIENCE 53
3.1 Introduction . . . 53
3.2 Definitions of Quality of Experience . . . 53
3.2.1 Quality of Experience Versus Quality of Service . . . 54
3.2.2 Quality of Experience Versus User Experience . . . 55
3.3 Factors Influencing Quality of Experience . . . 56
3.3.1 System Influence Factors . . . 56
3.3.2 Context Influence Factors . . . 58
3.3.3 Human Influence Factors . . . 60
3.3.4 Summary . . . 62
3.4 Features of Quality of Experience . . . 63
3.5 Methods for Multimedia Quality of Experience Evaluation . . . 64
3.5.1 Quantitative Methods . . . 65
3.5.2 Qualitative Methods . . . 66
3.5.3 Mixed Methods . . . 67
3.6 Discussion . . . 68
3.7 Summary . . . 68
I SYSTEM 71
4 A SUBJECTIVE COMPARISON OF DISCRETE SURROUND SOUND AND SOUNDBAR TECHNOLOGY 73
4.1 Introduction . . . 73
4.2 Methodology . . . 76

4.2.1 Structure . . . 76
4.2.2 Attribute Elicitation Introduction . . . 77
4.2.3 Familiarisation . . . 77
4.2.4 Preference Rating and Attribute Elicitation . . . 78
4.2.5 Attribute Refinement . . . 80
4.2.6 Attribute Rating . . . 80
4.3 Experimental Design and Setup . . . 81
4.3.1 Reproduction Systems . . . 81
4.3.2 Stimuli . . . 82
4.3.3 Participants . . . 84
4.3.4 Room . . . 84
4.3.5 Setup . . . 85
4.3.6 Calibration . . . 86
4.3.7 Administration . . . 87
4.3.8 Data Collection . . . 89
4.4 Results . . . 89
4.4.1 Participant Reliability . . . 89
4.4.2 Preference Ratings . . . 90
4.4.3 Sensory Profiling . . . 93
4.5 Discussion . . . 98
4.6 Summary . . . 101

II CONTEXT 103

5 ENVIRONMENTAL NOISE AND BACKGROUND-FOREGROUND AUDIO BALANCE 105
5.1 Introduction . . . 105
5.1.1 Audio Adaptation for Mobile Listening . . . 107
5.1.2 Background/Foreground Object Distinction . . . 108
5.2 Experimental Design: Study I . . . 109
5.2.1 Environmental Noise . . . 109
5.2.2 Audio Excerpts . . . 110
5.2.3 Reproduction Methods . . . 111
5.2.4 Setup . . . 112
5.2.5 Procedure . . . 114
5.2.6 Data Collection . . . 115
5.2.7 Participants . . . 115

5.3 Results: Study I . . . 116
5.3.1 Outlier Detection . . . 116
5.3.2 Participant Consistency . . . 116
5.3.3 Linear Mixed Model Analysis . . . 117
5.3.4 Cluster Analysis . . . 119
5.3.5 Discussion . . . 121
5.4 Experimental Design: Study II . . . 121
5.4.1 Environmental Noise . . . 122
5.4.2 Audio Excerpts . . . 122
5.4.3 Setup . . . 122
5.4.4 Procedure . . . 123
5.4.5 Data Collection . . . 124
5.4.6 Participants . . . 124
5.5 Results: Study II . . . 124
5.5.1 Level . . . 125
5.5.2 Ratio . . . 126
5.5.3 Cluster Analysis . . . 126
5.5.4 Nature of Ratio Adjustments . . . 127
5.5.5 Semi-Structured Interviews . . . 128
5.5.6 Discussion . . . 128
5.6 Experimental Design: Study III . . . 129
5.6.1 Audio Excerpts . . . 129
5.6.2 Procedure . . . 130
5.6.3 Data Collection . . . 131
5.6.4 Participants . . . 132
5.7 Results: Study III . . . 132
5.8 Qualitative Analysis . . . 134
5.8.1 Increased Background . . . 135
5.8.2 Increased Foreground . . . 136
5.8.3 No Change . . . 137
5.9 Discussion . . . 137
5.10 Summary . . . 140

III HUMAN 141

6 THE ROLE OF HUMAN FACTORS ON OVERALL LISTENING EXPERIENCE 143
6.1 Introduction . . . 143

6.2 Experimental Design . . . 144
6.2.1 Psychographic Data Collection . . . 145
6.2.2 Stimuli . . . 148
6.2.3 Procedure of Listening Sessions . . . 150
6.2.4 Data Collection . . . 152
6.2.5 Participants . . . 152
6.3 Results . . . 153
6.3.1 Participant Reliability . . . 153
6.3.2 Psychographic Data . . . 154
6.3.3 OLE Analysis: Part I . . . 157
6.3.4 Analysis of Listener Type . . . 160
6.3.5 OLE Analysis: Part II . . . 164
6.3.6 Interaction Between Psychographic Variables and Listener Type . . . 165
6.3.7 Complementary Analysis . . . 169
6.4 Discussion . . . 170
6.4.1 OLE of Binaural Audio . . . 170
6.4.2 Human Factors and OLE . . . 172
6.5 Summary . . . 173

7 DISCUSSION AND CONCLUSIONS 175
7.1 The Literature . . . 175
7.2 Part I: System . . . 176
7.2.1 Summary . . . 176
7.2.2 Discussion . . . 177
7.3 Part II: Context . . . 178
7.3.1 Summary . . . 178
7.3.2 Discussion . . . 178
7.4 Part III: Human . . . 180
7.4.1 Summary . . . 180
7.4.2 Discussion . . . 181
7.5 Overall Contribution . . . 181
7.6 Further Work . . . 183
7.7 Concluding Remarks . . . 184

REFERENCES 187

LIST OF FIGURES

Figure 1.1 Factors influencing QoE can be grouped into human, system and context influence factors. . . . 3
Figure 2.1 Traditional versus object-based broadcasting. . . . 16
Figure 2.2 Quality-formation process. . . . 21
Figure 2.3 Filter model. . . . 24
Figure 2.4 The relationship between validity and type of test. . . . 26
Figure 2.5 ITU-R five-grade continuous impairment scale. . . . 29
Figure 2.6 ITU-R continuous quality scale (CQS). . . . 30
Figure 2.7 Rating scales used in ITU-T P.800. . . . 31
Figure 2.8 Nine-point hedonic scale. . . . 33
Figure 2.9 Letwoski's MURAL. . . . 35
Figure 2.10 Relations between total audio quality and its subsets and attributes. . . . 36
Figure 2.11 Overview of the descriptive analysis process. . . . 38
Figure 2.12 Sound wheel for reproduced sound. . . . 43
Figure 3.1 Factors influencing QoE can be grouped into human, system and context influence factors. . . . 57
Figure 4.1 System: Overview of the original OPQ structure. . . . 76
Figure 4.2 System: Overview of the adapted OPQ structure. . . . 77
Figure 4.3 System: Familiarisation interface. . . . 78
Figure 4.4 System: Acoustic room measurements. . . . 85
Figure 4.5 System: Room dimensions and experimental setup. . . . 86
Figure 4.6 System: Preference rating and attribute rating user interfaces. . . . 88
Figure 4.7 System: Circular error rates. . . . 90
Figure 4.8 System: Marginal mean scaled preference scores. . . . 93
Figure 4.9 System: Word cloud of refined elicited attributes. . . . 94
Figure 4.10 System: PCA correlation loadings with attributes in the perceptual space. . . . 95
Figure 4.11 System: Objects and participants' preferences in the perceptual space. . . . 96

Figure 5.1 Context I: Environmental noise spectrograms. . . . 110
Figure 5.2 Context I: A comparison of attenuation functions from various models of headphones. . . . 113
Figure 5.3 Context I: Graphical user interface. . . . 114
Figure 5.4 Context I: Mean level and ratio with respect to environmental noise. . . . 118
Figure 5.5 Context I: Mean ratio for all conditions. . . . 119
Figure 5.6 Context I: Mean ratio with respect to environmental noise and cluster membership. . . . 120
Figure 5.7 Context II: Graphical user interface. . . . 123
Figure 5.8 Context II: Participant consistency. . . . 125
Figure 5.9 Context II: Mean level and ratio with respect to environmental noise. . . . 126
Figure 5.10 Context II: Mean ratio with respect to environmental noise and cluster membership. . . . 127
Figure 5.11 Context II: Component level vs. ratio. . . . 128
Figure 5.12 Context III: Graphical user interface. . . . 131
Figure 5.13 Context III: Histograms of chosen mix counts. . . . 132
Figure 5.14 Context III: Dot plots of the mean difference between without noise choices to with noise choices for each participant. . . . 133
Figure 5.15 Context: Main themes identified from thematic analysis. . . . 135
Figure 6.1 Human: User interface for OLE ratings. . . . 151
Figure 6.2 Human: BIR variance. . . . 153
Figure 6.3 Human: Distribution of data from demographic related psychographic questions. . . . 155
Figure 6.4 Human: Distribution of data from experience related psychographic questions. . . . 156
Figure 6.5 Human: Distribution of data from attitude related psychographic questions. . . . 157
Figure 6.6 Human: Histogram of all basic item ratings. . . . 158
Figure 6.7 Human: Relative frequencies of item ratings grouped by processing. . . . 158
Figure 6.8 Human: Colour map of average item ratings. . . . 159
Figure 6.9 Human: Relative frequencies of item ratings grouped by processing and content group. . . . 160

Figure 6.10 Human: Kendall's rank correlations between item rating and the two variables total quality level ($\tau_{IR,Q}$) and basic item rating ($\tau_{IR,BIR}$). 163
Figure 6.11 Human: Kendall's rank correlations between item rating and the two variables timbral quality level ($\tau_{IR,T}$) and total quality level ($\tau_{IR,Q}$) (Figure a) and the two variables spatial quality level ($\tau_{IR,S}$) and total quality level ($\tau_{IR,Q}$) (Figure b). 163
Figure 6.12 Human: Kendall's rank correlations between item rating and the two variables timbral quality level ($\tau_{IR,T}$) and spatial quality level ($\tau_{IR,S}$). 164
Figure 6.13 Human: Relative frequencies of item ratings grouped by processing, content and listener group. 165
Figure 6.14 Human: Correlations between measures of listener type and the psychographic variable competence with regression lines. 168
Figure 6.15 Human: Normalized item ratings (averaged over content) with respect to both overall quality level and competence group. . . 170

LIST OF TABLES

Table 1.1 Overview of IFs considered in each study. 6
Table 2.1 Synopsis of the four identified conceptual layers of sound quality. 22
Table 2.2 Assessor categorisation terminology. 27
Table 2.3 Examples of spatial and timbral attributes used to describe auditory events in the context of sound reproduction systems. . . 41
Table 3.1 Overview and examples of the various forms of influence factors. 62
Table 4.1 System: Overview of content items. 83
Table 4.2 System: Within-subject factor results from the mixed ANOVA model. 92
Table 4.3 System: Variability of principal components. 94
Table 5.1 Context I: Descriptions of environmental noises and audio items. 111
Table 5.2 Context I: Significant type III fixed effects for dependent variables level and ratio. 117
Table 6.1 Human: Overview of content items. 149
Table 6.2 Human: Results from Wilcoxon signed-rank tests on OLE data. 159
Table 6.3 Human: Kendall's τ correlation and significance between psychographic variables and measures of listener type. 166

1

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

The reproduction of sound is ubiquitous in modern life. From hearing aids to multichannel loudspeaker setups in cinemas, the forms and purposes of audio reproduction technology are incredibly varied. The experience provided by reproduced sound is, in part, a product of the technology associated with its capture and reproduction. Since the invention of the phonograph by Edison in 1877, audio reproduction technology has been continually evolving so as to provide more meaningful experiences to its users. This evolution is of course ongoing and today the next generation of audio reproduction technology offers possibilities of new and exciting experiences. Immersive audio reproduction by means of multichannel loudspeaker and binaural headphone setups offer a progression from the typical two-channel stereophonic setups to provide immersive, 3D experiences. Furthermore, object-based audio offers a progression from the typical channel-based delivery formats to provide greater personalisation, interaction and adaption of content to better suit the system, context and user.

Audio is an enhancing feature of many forms of media and therefore these experiential advances provided by next generation audio are not just relevant to audio-only consumption. Games, websites, online media, television, film, digital installations and virtual reality are but a few media forms that could be impacted by developments in next generation audio. Such media technology is also evolving to provide more immersive, personalised and mobile experiences in general, and it is therefore crucial for the audio aspects of these media forms to reflect this. With these media forms serving a range of purposes including to entertain, educate and inform, the potential roles and applications of next generation audio technology are broad, as are the associated experiences.

Consistent throughout the development of audio reproduction technology, and indeed electronic media technology in general, has been the need to evaluate its quality; it is necessary to understand how listeners perceive new technology in order to drive future developments. Subjective evaluation methods are typically used for this task

and a range of such methods are available in the literature. Perhaps the most commonly used subjective evaluation methods are those outlined by the International Telecommunication Union (ITU), such as recommendations ITU-R BS.1116 (ITU-R, 2015b) and ITU-R BS.1534 (ITU-R, 2015c). With these methods, experienced listeners in highly controlled laboratory settings are typically used and are intended to act as reliable quality meters. These standardised methods are evidently very suitable for certain applications. They are not, however, necessarily the most appropriate evaluation tool for all situations. For instance, in some situations it is desirable to gain insight into technology by eliciting attributes and preferences related to the provided experience; a process that is not represented in these standardised methods, which focus on quantitative assessment by means of global 'quality' scales. As such, a range of other quality evaluation techniques are also commonly developed and used.

In the field of multimedia[1] quality evaluation, the concept of quality of experience (QoE) has gained traction over the last few decades as an alternative to the more traditional quality of service (QoS) mindset. Whereas QoS is technology-centric and relates to service performance, QoE takes a more user-centric approach to quality evaluation and takes into account the person's context, personality and current state. More specifically, QoE can be defined as

> "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 fulfillment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the person's context, personality and current state"

(Möller and Raake, 2014, p. 19). A range of factors influence QoE and these can be grouped into system, context and human influence factors (IFs) (Le Callet, Möller, and Perkis, 2013). System IFs include factors related to aspects such as media capture, coding, transmission, storage, rendering, display and communication of information from content production to user, context IFs relate to properties of the user's environment and humans IFs relate to any variant or invariant property or characteristic of a human user. These groups of IFs often overlap and together have a mutual impact on QoE, as portrayed in Figure 1.1. Although the development of QoE is associated with multimedia evaluation (for instance audiovisual or web services), it is also ap-

1 Although the term "multimedia" can mean different things to different people, in general, multimedia includes a combination of content forms such as text, audio, still images, animation, video, and interactive content (Agnew and Kellerman, 2008).

Figure 1.1: Factors influencing QoE can be grouped into human, system and context influence factors. As represented in the figure, these groups of influence factors often overlap and together have a mutual impact on QoE. Adapted from (Möller and Raake, 2014, p. 57).

plicable for the evaluation of audio technology. As next generation audio technology may provide immersive experiences that can be adapted to the needs of the user and environment, it could be argued that a QoE mindset that considers system, context and human influence factors should be taken for its evaluation. Such an approach is in contrast to the commonly used standardised methods mentioned above, which have more in common with QoS, technology-centric approaches that focus only on system IFs.

Prior to commencing this project, the quality of experience of next generation audio was still a relatively unexplored area. It is beneficial for both consumers and developers of technology, however, to increase our insight of this. A greater quality of experience would provide more meaningful experiences to users of the technology and this could be achieved if developers have the tools to assess what constitutes a good QoE. Furthermore, next generation audio technology will likely have a greater impact and be able to provide more meaningful experiences if it is created and evaluated with the increasingly effective models of quality that QoE research brings.

It is worth noting that parallels can be drawn between the research areas of QoE and human-computer interaction (HCI), an interdisciplinary field that emerged from the research area of computer science. As the name suggests, HCI is concerned with

investigating and designing interactions between users and computer technology and, as such, is inherently user-centric. This is reflected in the "user-centered design" approach that is often advocated in HCI. Preece, Rogers, and Sharp, (2002, p. 286) suggest five principles that clarify the meaning of a user-centered approach and these include capturing and designing for user characteristics, and studying user behaviour and context of use so as to design to support them. QoE and HCI therefore both share the concern for user and context factors. It should also be noted that the design and development of next generation audio technology in general can be considered in the scope of HCI. One aspect of HCI is the design and delivery of new experiences through media technology, for example see (Churchill and Bardzell, 2007). This is related to the broader emerging concept in HCI of designing for experience, or "experience-centered design" (McCarthy and Wright, 2004; Wright and McCarthy, 2010), which can be considered as "a humanistic approach to designing digital technologies and media that enhance lived experience" (Wright and McCarthy, 2010, p. vi). In designing next generation audio technology, one is ultimately designing for new experiences through media technology. The design and evaluation of next generation audio is therefore not just a relevant topic to the research fields of traditional audio quality evaluation and QoE, but also HCI.

With the above background and motivation in mind, we are able to set the scope for the following thesis.

1.2 AIMS AND OBJECTIVES

In general, the aim of this research was to investigate the quality of experience provided by next generation audio technology. More specifically, the following two aims were addressed:

I. To explore the role of system, context and human influence factors on the quality of experience of next generation audio.

II. To investigate suitable methods and approaches for the evaluation of the quality of experience of next generation audio, with respect to its various influence factors.

These aims were explored by means of three case studies, each focussing on a particular category of influence factor and a particular form of next generation audio

technology. The three case studies had the following main specific objectives, the rationale of which are discussed more thoroughly in the forthcoming chapters:

I. System: To subjectively compare soundbar technology with discrete surround sound technology.

II. Context: To investigate whether environmental noise influences preferred audio object balance.

III. Human: To investigate the role of human influence factors on overall listening experience.

The first of these objectives primarily deals with the role of system IFs on the QoE of next generation audio. The study related to this objective is typical of those found in the field of audio quality evaluation, where system IFs are predominantly studied. Additionally, the experiment related to this objective considered human IFs in the form of a comparison between experienced and naïve listeners.

The second of these objectives deals with how next generation audio can be utilised to improve QoE in relation to context IFs. Specifically, advantages of object-based audio are studied with respect to the context influence factor of environmental noise. System IFs are also considered by comparing the effect of different reproduction methods. Three experimental studies are presented regarding this objective.

The third objective focusses on how the QoE of next generation audio could be influenced by various human IFs. More specifically, the role of human IFs such as attitudes, demographics and experiences are related to the influence of technical audio quality on overall listening experience. Additionally, system IFs were considered by evaluating the QoE provided by binaural audio.

It is therefore the case that the main categories of IFs are not investigated in isolation. As the three groups of influence factors have a mutual impact on QoE (Figure 1.1), it can be useful to study multiple IFs concurrently. Thus, each of the above case studies deals with multiple IFs. An overview of the influence factors considered in each study is given in Table 1.1.

As well as reflecting on the specific objectives above, this thesis also reflects on the effectiveness of each case study as an exemplar of its class of IF. Moreover, methodological approaches and results from the specific case studies are related to the investigation of IFs more generally, thus broadening the insight gained from the individual case studies.

Table 1.1: Overview of IFs considered in each study. Those in italic type are related to secondary objectives.

	System IFs	Context IFs	Human IFs
Study I - System	Reproduction methods		*Experience*
Study II - Context	*Reproduction methods*	Environmental noise	
Study III - Context		Environmental noise	
Study IV - Context		Environmental noise	
Study V - Human	*Reproduction methods*		Psychographic variables

1.3 ORIGINAL CONTRIBUTIONS

On a high-level, the research presented here contributes the thesis that to effectively evaluate the perceived quality of next generation audio, a QoE mindset should be taken that considers system, context and human influence factors. Five empirical studies are presented that highlight ways in which these influence factors are important for the assessment of next generation audio. Moreover, suitable methods are presented to study such influence factors. These methods are not just limited to the studies outlined in this thesis, but can be expanded to investigate other aspects of human, context and system IFs.

More specific contributions to the fields of audio quality evaluation and quality of experience were made and these are summarised in the following points.

- A subjective comparison of soundbar and discrete surround sound technology has been undertaken. This provides an original contribution as it is the first thorough perceptual evaluation of soundbar technology.
- The method Open Profiling of Quality was applied to the comparison of audio reproduction systems and was thus introduced to the field of audio quality evaluation. This contribution provides an additional tool by which practitioners can relate overall preference with sensory percepts, so that insight can be gained into the experience provided by audio systems.

- An understanding was gained of both how and why environmental noise influences preferred audio object balance. The studies presented in this thesis are the first to empirically investigate this matter and contribute to the understanding of how object-based audio can be used to improve QoE.
- The influence of technical audio quality on overall listening experience was related to a range of psychographic variables. This provides valuable insight into why human influence factors need to be addressed in order to optimise the QoE of next generation audio technology. Furthermore, the method outlined in this paper to study the link between psychographic variables and audio quality contributes a useful tool for future studies in this area.
- An assessment of the impact of binaural processing on overall listening experience has been presented. This is the first study to evaluate binaural audio with the metric overall listening experience and contributes to the understanding of the relative importance of binaural processing on QoE.

1.3.1 *List of Publications*

At the time of writing, the following publications have arisen as a result of the work described in this thesis.

Peer-Reviewed

I. Walton, T., Evans, M., Kirk, D. and Melchior, F (2016). "A subjective comparison of discrete surround sound and soundbar technology by using mixed methods." In: *Audio Engineering Society Convention 140*.[1]

II. Walton, T., Evans, M., Kirk, D. and Melchior, F (2016). "Does environmental noise influence preference of background-foreground audio balance?" In: *Audio Engineering Society Convention 141*.[2]

1 Also available as: Walton, T., Evans, M., Melchior, F. and Kirk, D. (2016). "A subjective comparison of discrete surround sound and soundbar technology by using mixed methods." In: *BBC Research & Development White Paper, WHP 320*

2 Also available as: Walton, T., Evans, M., Melchior, F. and Kirk, D. (2016). "Does environmental noise influence preference of background-foreground audio balance?" In: *BBC Research & Development White Paper, WHP 325*

III. Walton, T., Evans, M., Melchior, F. and Kirk, D (2017). "Combining preference ratings with sensory profiling for the comparison of audio reproduction systems." In: *Audio Engineering Society Convention 142*.[1]

IV. Walton, T. and Evans, M. (2018) "The role of human influence factors on overall listening experience". In: *Quality and User Experience*. Advance online publication. doi: 10.1007/s41233-017-0015-4.

V. Walton, T., Evans, M., Kirk, D. and Melchior, F. (2018) "Exploring object-based content adaptation for mobile audio." In: *Personal and Ubiquitous Computing*. Advance online publication. doi: 10.1007/s00779-018-1125-6.

Non Peer-Reviewed

VI. Walton, T. (2017). "The overall listening experience of binaural audio." In: *Proceedings of the 4th International Conference on Spatial Audio (ICSA)*.

The co-authors listed in the above publications supervised the research that forms the basis for this thesis. They did not contribute to the writing of the material and therefore sections from these publications are reproduced in this thesis. Publications I and III are related to the system IF study and material from these is found in Chapter 4, publications II and V are related to the context IF studies and material from these is found in Chapter 5 and publications IV and VI are related to the human IF study and material from these is found in Chapter 6.

1.3.2 *Data Access*

Data supporting this thesis is openly available under an 'Open Data Commons Open Database License'. Additional metadata are available at:

http://dx.doi.org/10.17634/154300-81.

Please contact Newcastle Research Data Service at rdm@ncl.ac.uk for access instructions.

1 Also available as: Walton, T., Evans, M., Melchior, F. and Kirk, D. (2017). "Combining preference ratings with sensory profiling for the comparison of audio reproduction systems." In: *BBC Research & Development White Paper, WHP 329*

1.4 STRUCTURE OF THE THESIS

The thesis is organised as follows. Chapter 2 describes related work in the field of audio technology and its evaluation. This chapter includes an overview of the history of audio reproduction technology including next generation audio, discussions on the term "quality" and its formation, an overview of methods used for the quality assessment of audio and a review of the state of the art of next generation audio evaluation.

Chapter 3 describes related work in the field of QoE. In this chapter, the concept of QoE is introduced with discussions on definitions of QoE, factors that influence QoE are presented, features of QoE are discussed and a brief overview of methods for the evaluation of QoE is given.

The following sections of the thesis are split into three parts: Part I - System, Part II - Context and Part III - Human. Each of these parts contain chapters of the empirical studies related to each of these categories of influence factors.

In Part I, Chapter 4 presents the first empirical study, in which the role of system IFs on the QoE of next generation audio are investigated. This chapter follows the structure of an introduction including specific related literature, methodology, experimental design and setup, results, discussion and summary.

In Part II, Chapter 5 presents the three studies related to how next generation audio can be utilised to improve QoE in relation to context IFs. After a general introduction, there are sections on experimental design and results for each of the studies, before sections on qualitative analysis, a discussion and summary consider the results as a whole.

In Part III, Chapter 6 presents the final empirical study, which investigates how the QoE of next generation audio could be influenced by various human IFs. Again, the chapter follows the structure of an introduction including specific related literature, experimental design, results, discussion and summary.

Finally, Chapter 7 concludes the thesis with discussions on the conducted research in relation to the aims presented in this chapter. The original contributions of the thesis are highlighted and discussed in conjunction with the limitations of this work and areas for further study.

2

AUDIO TECHNOLOGY AND ITS EVALUATION

2.1 INTRODUCTION

In order to be in a position to evaluate the quality of experience of next generation audio, it is first necessary to have a grounding in the field of audio reproduction technology and its evaluation. It is this which is the topic of the current chapter. We begin this chapter with an overview of the progression of audio technology from mono to the current next generation technologies, Section 2.2. This is followed by considerations on the term "quality" in Section 2.3. The majority of this chapter deals with audio quality evaluation; Section 2.4 presents principles of audio quality evaluation, Section 2.5 presents commonly used global judgment evaluation methods, Section 2.6 presents a range of attribute-based evaluation methods, Section 2.7 discusses methods that combine global and attribute-based aspects and Section 2.8 briefly discusses sound quality models. The state of the art of next generation audio evaluation is presented in Section 2.9, followed by a discussion in Section 2.10 and a summary in Section 2.11.

2.2 FROM MONO TO NEXT GENERATION

In this section a history of audio reproduction technology from the earliest monophonic systems to next generation systems is presented. A key feature of next generation audio is the immersive experiences that it can potentially deliver due to advances in spatial reproduction. Whilst timbral aspects of audio reproduction have also improved greatly since the earliest systems, it is advances in spatial aspects that will help define systems in the coming years. It is due to this that this section has a focus on the spatial aspects of audio reproduction history.

2.2.1 *Reproduction in the Horizontal Plane*

One of the fundamental ways in which we make sense of our environment is through the process of hearing. Natural sounds typically contain cues in all three dimensions (width, height, depth) and can be perceived as possessing a certain size in space. The ability for humans to process these three-dimensional cues plays a major role in the formation of spatial awareness and hence helps to form an understanding of the environment around us (Rumsey, 2001). Reproduced sound historically has not replicated the three-dimensional cues necessary to form an accurate spatial representation of an environment. The earliest sound reproduction systems, gramophones and monographs from the late 1800s and early 1900s, were monophonic (i.e. single-channel). With monophonic reproduction, no directional spatial cues are reproduced - only distance cues provided by reverberation are replicated.

An advancement of this came in the form of two-channel stereophonic reproduction. The first known example of a stereophonic transmission of music is documented as occurring in 1881 during an experiment by Clément Ader (Hertz, 1981). Work by Alan Blumlein in the 1930s (Blumlein, 1931) is regarded as being the beginning of modern stereophonic technology, although it was not until the 1950s that these techniques were introduced commercially. More than half a century on, two-channel stereophonic reproduction remains the standard for the majority of broadcast material. With two-channel stereophonic reproduction it is possible to create the effect of a sound source originating from any point on a line between two loudspeakers. Positioning of a sound source in a stereophonic reproduction can be achieved by adjusting the source's relative amplitude in each speaker (amplitude panning), relative delay in each speaker (time-based panning), or both. These techniques therefore produce interaural time differences (ITDs) and interaural level differences (ILDs) at the listening position, which are the hearing cues responsible for the detection of source position in the horizontal plane.

Multichannel horizontal plane systems such as 5.1 surround sound and the lesser used 7.1 surround sound (ITU-R, 2012a) are a continuation of two-channel stereophonic techniques. Such layouts place extra channels around the listener with the aim of generating a more enveloping experience. However, it should be noted that such systems are not intended to be able to reproduce accurate 360° image localisation (Rumsey, 2001).

2.2.2 *Immersive Audio*

The stereophonic techniques mentioned above are all confined to reproduction on a horizontal plane, i.e. they produce two-dimensional sound scenes. This is typically not how we hear sound in the real-world, as discussed earlier in this section. Immersive audio, also referred to as 3D audio, adds the dimension of height to the listening experience. Despite there being attempts to popularise immersive audio as early as the 1940s, it is not until recently that immersive audio is becoming a viable option for commercial installations in the home and cinema (Rumsey, 2015).

A range of multichannel speaker layouts have been proposed to deliver immersive audio, including 9.1, 10.2 and 22.2 (ITU-R, 2015a), among others. Whilst it is not yet agreed which speaker layout is optimal, these systems share the aim of enriching the listening experience by including elevated loudspeakers. Traditional two-dimensional stereophony (time and/or level panning) can be extended into the third dimension for use with such immersive setups to create virtual sources, for example with vector base amplitude panning (VBAP) (Pulkki, 1997). Hacihabiboglu et al., (2017) present an overview of such perceptually motivated reproduction methods. These techniques have the advantage of being computationally simple, although spatial fidelity is compromised outside of a central sweet spot (Spors et al., 2013).

Whereas the aim of traditional stereophonic systems is to create a plausible auditory impression based on panning laws, sound field synthesis techniques aim to accurately reconstruct a physical sound field over an extended listening area. Higher-Order Ambisonics (HOA) is such a technique, the basic concepts of which date back to the 1970s (Gerzon, 1973). HOA is a reproduction method that works by capturing a loudspeaker-independent three-dimensional representation of a sound field (based on a spherical harmonic decomposition) and decoding this to a loudspeaker array. Another such technique is Wave Field Synthesis (WFS) (Berkhout, Vries, and Vogel, 1993). One key property of sound field synthesis techniques is that they allow the creation of sources between the loudspeaker radius and listener (focussed sources), as well as beyond the loudspeaker radius. In comparison, with stereophonic techniques it is not possible to create focussed sources and reverberation effects need to be used to simulate sources beyond the loudspeaker radius. A limiting feature of sound field synthesis techniques is that for accurate sound field representation, a large number of loudspeakers are necessary.

So far, this short review of immersive audio has been focussed on loudspeaker reproduction. However, perhaps the most accessible form of immersive audio is achieved through headphone technology. Binaural audio has the aim of delivering independent signals to the two ears that contain the natural human binaural cues that are necessary for proper spatial localisation. The shape and size of ones ear, head and, to a lesser degree, torso, means that the frequency response of an incoming sound signal is modified (or filtered) in a certain way depending on the direction of the source. As this process is anatomy dependent, there are significant differences in head-related transfer functions (HRTFs) between individuals. It is these cues that binaural audio utilises. Experiments in binaural audio date back to the 19th century (Moncel, 1881), but it was around the 1930s that research into the topic developed (Blumlein, 1931). See (Paul, 2009) for a full historical review.

The simplest method to achieve binaural audio is to record an acoustic scene with microphones at the eardrums of a listener or dummy head and then reproduce the recorded signals via headphones, thus preserving the spatial information of the scene. There are several key limitations to this method of binaural reproduction including the inability to personalise the HRTFs used, the inability to alter the sound depending on the head movements of a listener and general impracticalities of the recording technique. An alternative approach, which has been made more viable due to increases in computer processing power, is to synthesise binaural signals by processing non-binaural audio with position dependent HRTF measurements. This allows for dynamic rendering as well as HRTF personalisation, both of which are considered to be important factors for an effective, externalised reproduction (Begault, Wenzel, and Anderson, 2001). Furthermore, through binaural synthesis existing multichannel material can be rendered to "virtual loudspeakers" with the aim of simulating the experience of listening to a multichannel loudspeaker setup in a room (Horbach et al., 1999).

2.2.3 *Object-Based Audio*

As spatial audio reproduction hardware advances, so too has the format with which audio content is delivered. Two-dimensional stereophonic setups traditionally use channel-based production techniques. In other words, signals directly relating to a certain loudspeaker layout are stored and transmitted, for example a left and a right signal in a stereo file corresponding to a left and right loudspeaker. Such a technique

assumes that the listener has the corresponding loudspeaker setup to the one used in the mixing of the audio content. As we have seen in the previous section, trends in audio are heading towards a higher and higher channel count with a variety of possible layouts. With a greater number of loudspeakers to arrange, it is unlikely that it will be practical for consumers to replicate the layout used in the mixing environment in their homes. Moreover, in a channel-based workflow multiple mixes would need to be made on the production end to satisfy the various immersive audio layouts and techniques.

A scene-based approach is one alternative to this. With a scene-based representation an entire sound field is stored as orthogonal basis functions, which can then be decoded to a given loudspeaker layout. Higher Order Ambisonics is one example of a scene-based technique.

Another approach is object-based audio. This is a method of representing sound as separate elements (or "objects") with corresponding temporal, positional and other/semantic metadata, so that the objects can be rendered with a large degree of flexibility at the user's end (Herre et al., 2015; Kim, 2014; Melchior, Churnside, and Spors, 2012). For example, instead of mixing a source to a certain loudspeaker channel, the source object is transmitted with positional metadata which the renderer on the user's end can use to reproduce the intended source position. The renderer that pieces together the objects is therefore able to account for the speaker setup so as to provide an optimal listening experience. Furthermore, one mix is able to accommodate for the various reproduction systems that might be used by the listener (e.g. stereo/binaural/22.2), thus reducing production costs. For a schematic comparison between traditional and object-based broadcasting concepts, see Figure 2.1.

Additional advantages of an object-based approach are the possibilities for greater personalisation, interaction and content adaptation (Armstrong et al., 2014; Parmentier, 2015). Such adaptation includes adaptation to suit the device or system (as discussed above), but also adaptation to suit the environment and adaptation to suit the user. For example, Mann et al., (2013) experimented with an object-based audio broadcast of a live football match where listeners were given audio feeds from opposite ends of the stadium together with a commentary feed and were able to mix the balance to suit their preference. Roughly three quarters of listeners preferred the object-based experience compared to traditional radio coverage. Other examples of object-based experiences include mix adaptation for hearing impaired listeners (Shirley and Oldfield, 2015), visual content adaptation to suit the user's profile (Evans et al., 2016) and

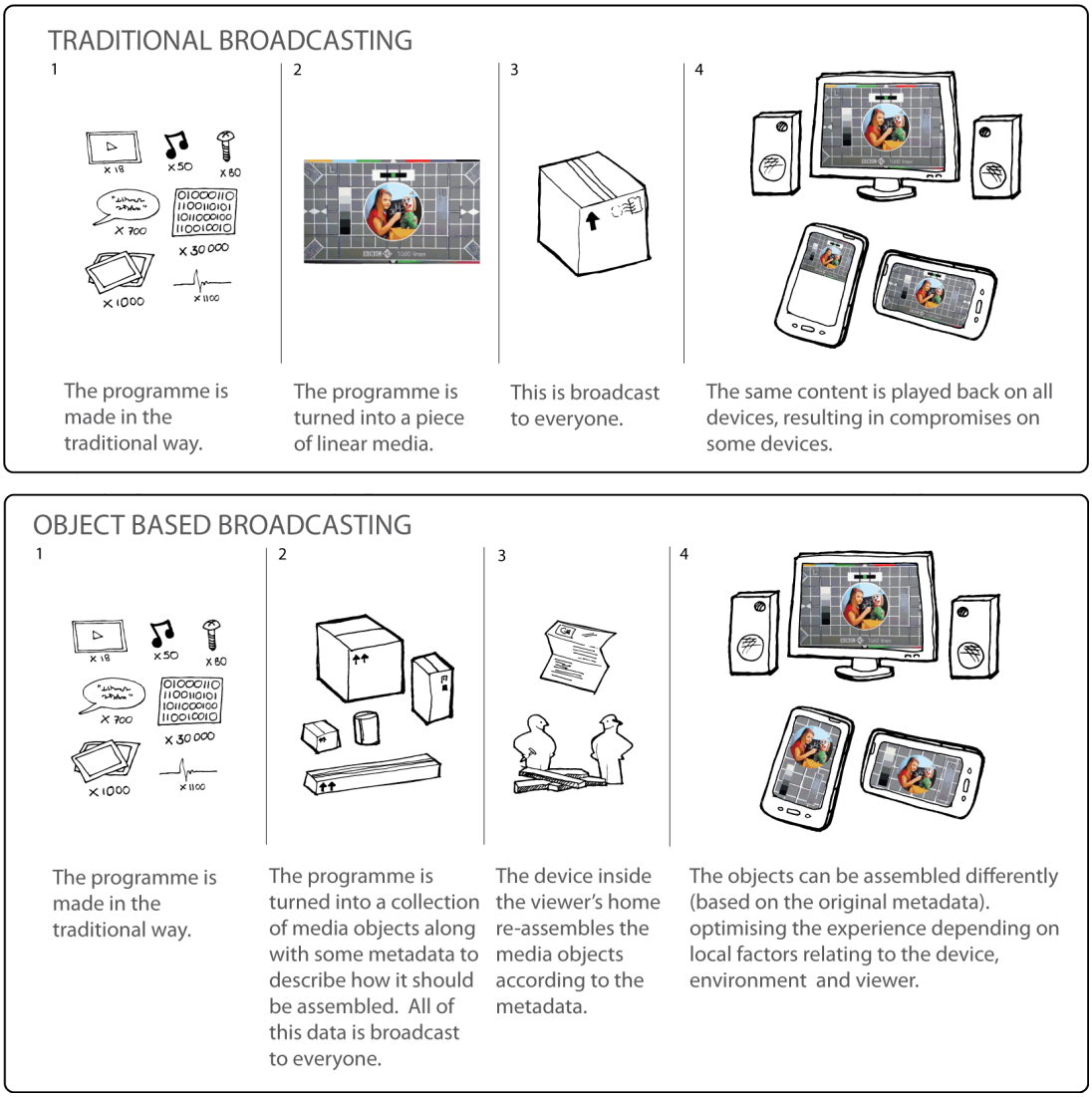


Figure 2.1: Traditional versus object-based broadcasting (BBC R&D, Jasmine Cox, 2013).

adapting the length of audio content to suit the user's requirements (Armstrong et al., 2014).

2.2.4 *Next Generation Audio: A Summary*

Rumsey, (2006) highlights the fact that the technical quality of reproduced sound in high-end systems in terms of factors such as a flat frequency response and low levels of distortion is becoming asymptotic to the ideal. Spatial quality, however, has some way to go before it reaches this point. This coupled with the fact that academic research trends in spatial audio of the last decade are finding their way into commercial application (Melchior, Churnside, and Spors, 2012), suggests that it is the spatial aspect of reproduced sound which will see the biggest developments in the coming years.

With a growing rise in the popularity of mobile devices, an increasing number of individuals are consuming audio over headphones. This, coupled with the rise in popularity of virtual reality (VR), means that binaural audio will likely be at the forefront for taking immersive audio from the research setting to the domestic.

Immersive systems with a greater channel count are now available, such as 9.1, 10.1 and 22.2, but it is unlikely that such systems will appeal to the majority of consumers in a domestic setting due to practicalities of correctly positioning a large number of loudspeakers. One device that could make spatial audio a more viable technology for domestic use is the soundbar. Such technology is generally advertised as being able to deliver a good spatial impression from a single enclosure of speakers. In recent years, soundbars have seen a large rise in interest (Zion Market Research, 2016) as such devices are possibly viewed as being simpler to setup, more convenient and more aesthetically pleasing than traditional discrete setups. Spatialisation from soundbars can be achieved in several ways. One method is to use beamforming with an array of speakers (Hooley, 2006) - the aim being to reflect sound around the room so that it appears as if there are speakers positioned around the listener. Another method is to take advantage of psychoacoustic phenomena to create a virtual surround effect.

Media device orchestration (Francombe et al., 2017a) is another technique that presents a more practical approach to immersive audio reproduction in a domestic environment. With this approach, media devices with loudspeakers that are already present in the room (such as mobile phones, tablets and laptops) are utilised to augment audio content in order to provide an immersive experience. This concept is still

in its infancy, although the development of internet of things and object-based audio technologies mean that this could be an increasingly relevant area of research in the coming years.

Accompanying possible improvements to the listening experience due to immersive audio techniques are possible improvements due to personalisation, adaptation and interactivity. Object-based audio allows for such improvements.

Considering the review of next generation audio in this section, it could be seen that the concept of next generation audio is defined by more than just emerging technological trends such as immersive multichannel reproduction systems and object-based technologies, but also by the associated experiences they provide. A key feature of next generation audio technologies is the aim to provide immersive, interactive and personalised experiences. The provision of these experiences could therefore be seen as a defining factor of next generation audio. Furthermore, associated behavioural trends could help define next generation audio technology, as well as shape how it develops. For instance, ubiquitous listening and the extensive use of audio devices with significant processing capabilities are two trends that will likely influence the development of next generation audio technology and thus help define it. It should be noted that such characteristics and trends are not unique to next generation audio. They are also key features of emerging media consumption technologies and interactive technologies more broadly and are therefore current topics of interest in HCI as a whole.

In order to be able to properly evaluate the experience provided by next generation audio technology, we need to have suitable evaluation methods and tools. The focus of the following sections is to provide an overview of assessment methods typically used for the evaluation of audio technology.

2.3 CONSIDERATIONS ON THE TERM "QUALITY"

Before discussions are presented on the topics of audio quality evaluation and quality of experience, it is important to consider what the term "quality" means in general and how quality judgements are formed. In this section some definitions of the term "quality" are given, a theoretical model for the quality-formation process is presented and a model of sound quality is introduced. It should be noted that in this section "quality" is discussed in a general sense, i.e. not specifically related to audio, unless

otherwise stated. These concepts on "quality" in general are however also applicable to the specific case of audio quality.

2.3.1 *Definitions of "Quality"*

An overview of some early considerations on, and the evolution of, the term "quality" are presented by Möller and Raake, (2014). They observe that over the last 15 years or so (at the time of publication) the definition of the term quality has radically changed and that this coincides with a general lack of understanding of the term.

A definition of quality from the year 2000 refers to the "totality of characteristics of an entity [...] that bear on its ability to satisfy stated or implied needs" (ISO, 2000). Möller and Raake relate this definition to what is now referred to as the "character" of an entity, as discussed by Blauert and Jekosch, (2003). The ISO definition of quality was updated in 2005 to read that quality is the "degree to which a set of inherent characteristics [...] fulfils requirements" where a characteristic is a "distinguishing feature" (ISO, 2005).

A definition given by Jekosch, (2005) states that quality is the "result of judgment of the perceived composition of an entity with respect to its desired composition". This is drawn upon in the Qualinet white paper on definitions of quality of experience to provide the following definition (Le Callet, Möller, and Perkis, 2013, p. 4):

> "[Quality] is the outcome of an individual's comparison and judgment process. It includes perception, reflection about the perception, and the description of the outcome. In contrast to definitions which see quality as 'qualitas', i.e. a set of inherent characteristics, we consider quality in terms of the evaluated excellence or goodness, of the degree of need fulfilment, and in terms of a 'quality event'." [1]

It is clear from the latter two definitions that quality can be seen to relate to an individual's point-of-view and that quality-formation involves a perception and a judgement process. It is this quality-formation process that will now be discussed.

1 "Definitions which see quality as 'qualitas"' refers to definitions of quality such as the first ISO definition presented above.

2.3.2 *The Quality-Formation Process*

A theoretical model for the quality-formation process is given by Jekosch, (2004). In this model the first part of the quality-formation process is based around the interaction between the physical and perceptual domains. An entity is described in the physical domain by "quality elements" and in the perceptual domain by "quality features". Quality elements and quality features interact in a multivariate way; every quality element may influence several quality features and every quality feature may be influenced by several quality elements. This multivariate process is said to be "neither necessarily linear and additive nor time-invariant" - a comment noting the complexity of the quality-formation process. Blauert, (2005) describes this interaction between the physical and perceptual domains as the bottom-up, signal-driven part of the quality judgement process (bottom-up processes are signal-driven and pre-attentive whereas top-down processes are driven by cognition).

A second section of the quality-formation process is the comparison of these external inputs (formed from the multivariate relationships between quality elements and quality features) and an internal reference of the perceiver. This internal reference reflects the temporal and contextual nature of the quality-formation process and is highly dependent on the task and application. This part of the quality-formation process is top-down, hypothesis-driven.

Figure 2.2 is a schematic diagram of the quality-formation process and shows how the external inputs and internal reference interact. It is seen that as well as temporal and contextual factors, memory of former experienced qualities can influence the reference path as indicated by the arrow from experienced quality to the reference path. The outcome of the quality-formation process is an experienced quality that is delimited in time, space and character (a quality event). As this event happens inside the human user, relevant information about the event can only be obtained on a descriptive level from the user (Le Callet, Möller, and Perkis, 2013). This quality-formation process can be seen to consist of a mental process of comparison and distance ratings between the two paths (Blauert and Jekosch, 2003).

2.3.3 *Sound Quality*

The specific case of sound quality is now considered. Blauert and Jekosch, (2003) identify that "as the references play such a paramount role in the quality-formation

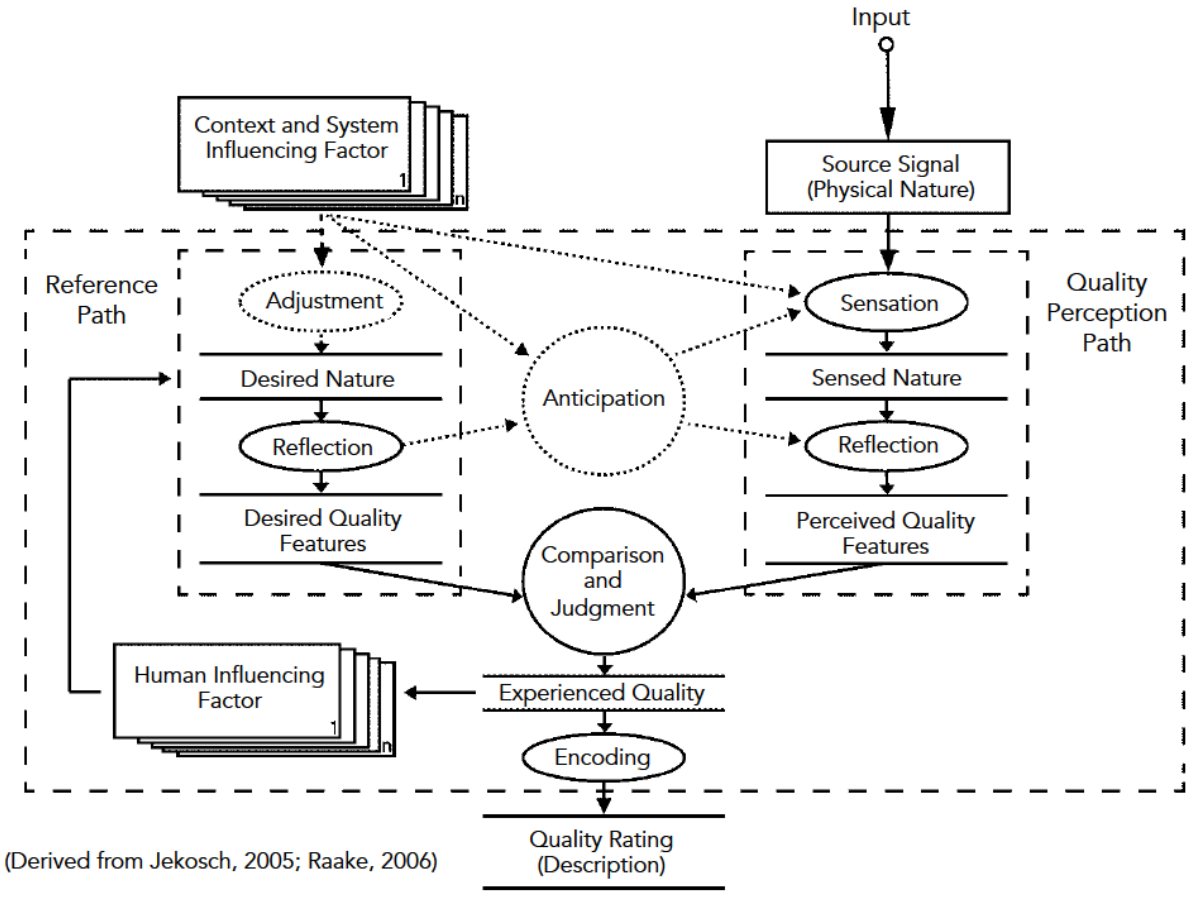


Figure 2.2: Quality-formation process (Le Callet, Möller, and Perkis, 2013).

Conceptual Aspect	*Examples of Issues*	*Suitable Measuring Methods*
Auditive Quality *Classical Psychoacoustics*	Perceptual properties such as loudness, roughness, sharpness, pitch, timbre, spaciousness	*Indirect scaling:* thresholds, difference limens, points of subjective equality *Direct scaling:* category scaling, ratio scaling, direct magnitude estimation
Aural-scene Quality *Perceptual Psychology*	Identification and localization of sounds in a mixture, speech intelligibility, audio perspective incl. distance cues, scenic arrangement, tonal balance, aural transparency	*Discretic:* semantic differential, multi-dimensional scaling. *Syncretic:* scaling of preference, suitability, and/or appropriateness, benchmarking against target sounds
Acoustic Quality *Physics*	Sound-pressure level, impulse response, transmissions function, reverberation time, sound-source position, lateral-energy fraction, inter-aural cross correlation	Instrumental measurements with physical equipment for the measurement of elasto-dynamic vibrations and waves, including appropriate signal processing
Aural-communication Quality *Communication Sciences*	Product-sound quality, comprehensibility, usability, content quality, immersion, assignment of meaning, dialogue quality	Psychological (cognitive) tests, particularly in realistic use cases, e.g., the product in use, the audience in concert, etc., questionnaires, dialogue tests, comprehension test, usability tests, market surveys

Table 2.1: Synopsis of the four identified conceptual layers of sound quality (Blauert and Jekosch, 2012).

process, but are not readily available for physical or psycho-acoustic measurement, they obviously pose a problem". It is with this in mind that they introduce the idea of classifying references from the amount of abstraction given in the reference characters. This results in a layer model of sound quality according to the amount of abstraction involved (Blauert and Jekosch, 2012). This model is based on the perceptionist's assumption that any- and everything that exists in the world is a precept linked to brain function. Four quality layers are identified (in order of increasing abstraction): auditive quality, aural-scene quality, acoustic quality and aural-communication quality. These layers are outlined in in Table 2.1. The layer of lowest abstraction is auditive quality. This layer is based around auditory events - precepts that exist at a specific time at a specific location in space, distinct by their characteristic properties. Such properties include loudness, pitch, timbre, roughness, sharpness, position and spatial extent. Classical psychoacoustics attempts to describe these characteristics without influence of cognitive interpretation, that is, by avoiding any abstraction. The next layer is aural-scene quality. In this layer object building and perceptual grouping is involved. Auditory effects such as the precedence effect, auditory-stream segregation, the cocktail-party effect and melody recognition can be found here. A significantly higher degree of abstraction is required by the next layer: acoustic quality. In this layer features of auditory events are compared with physical quantities which is said to require a conceptual model of the world. This layer includes measures such as sound pressure level, reverberation time and inter-aural cross correlation. The layer of highest abstraction is aural-communication quality. At this level of abstraction audi-

tory events are conceived as carriers of signs which refer to feelings, things of concepts. Here measures such as product-sound quality, usability and immersion are found.

2.4 AUDIO QUALITY EVALUATION PRINCIPLES

A fundamental process in the advancement of audio technology is quality evaluation. It is crucial to understand how listeners perceive new technology in order to drive future developments. Physical measurements of audio signals in the acoustic and electrical domains offer some insight into audio quality, although their use is limited as they do not directly relate to how humans perceive sound. Audio is multidimensional in nature and, as such, in order to predict audio quality from physical measures complex quality models are needed. Such models are ultimately desirable due to reliability, repeatability and lower resource requirements. However, the complex, multidimensional nature of quality means that their development depends on first fully understanding the relationship between sensory percepts, overall experience and the corresponding physical measures. Perceptual evaluation on the other hand directly measures perceived quality by asking participants to quantify their experience, often by means of listening tests. It is this form of quality evaluation that is the focus of the following sections. Before specific examples of listening test methods are given in sections 2.5, 2.6 and 2.7, discussions are first presented on possible types of measurement, validity and assessor categorisation.

2.4.1 *Physical, Perceptual and Affective Measurement*

As mentioned above, audio can be measured both in the physical and perceptual domains. Furthermore, perceptual evaluation can take various forms depending on the purpose of the experiment. Figure 2.3 presents the filter model, originally introduced by Pedersen and Fog, (1998) and modified by Bech and Zacharov, (2006), which illustrates the relationships between these different forms of audio measurement. At the input of the model we find a complex acoustic stimulus. Such physical stimuli are characterised by physical measurements, for example frequency and sound pressure level, and are measured by technical measuring equipment, such as sound pressure level meters.

The first filter - "the senses" - transforms these physical stimuli into perceived stimuli or "auditory events". Stimuli at this stage of the model are in the mind of

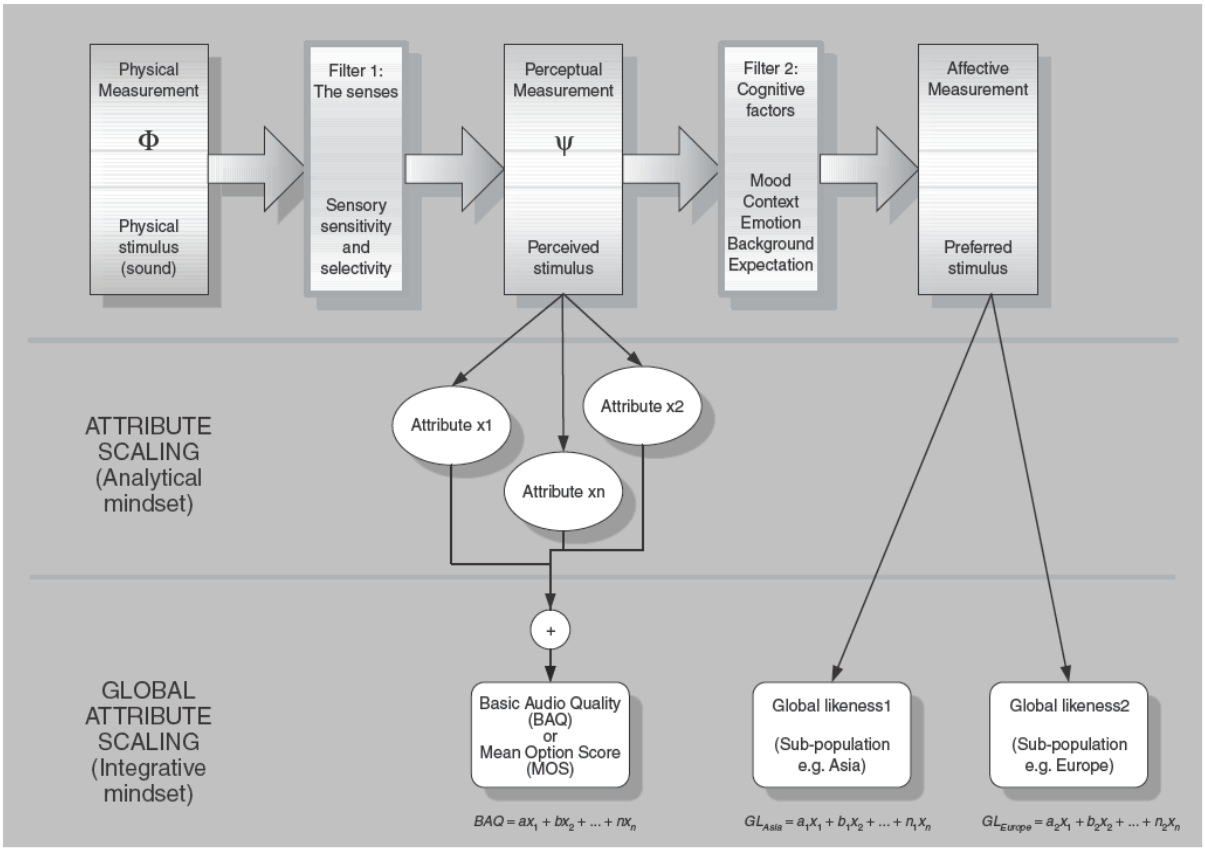


Figure 2.3: Filter model as presented in (Bech and Zacharov, 2006).

the listener and can be composed of a number of individual auditory attributes, for example pitch (perceived frequency) and loudness (perceived magnitude). Auditory events are characterised by the perceptual measurement of relevant attributes, often by means of expert panels of listeners. Research questions associated with this stage of the model could be related to the quantification of a single attribute, or if a complex stimulus such as music is used, could be related to the quantification of multiple attributes. Moreover, an integrative mindset at this stage could lead to the combination of attributes into a global quality score such as Basic Audio Quality (BAQ).

The second filter - 'cognitive factors' - transforms perceptual measures into affective measures. Such measures are influenced by factors such as mood, context, emotion, background and expectation and are associated with research questions related to, for example, preference, annoyance and acceptance. Typically, these types of measurements are conducted by consumer panels, i.e. the end users of a product.

Further considerations on the filter model are made by Pedersen, (2009). It is highlighted that linking physical measures with perceptual measures is the aim of perceptual models and linking perceptual measures with affective measures is the aim of preference mapping.

2.4.2 *Validity*

When assessing the applicability of different evaluation techniques it is necessary to consider the issue of validity, the topic of which is related to the concepts presented in the filter model. A distinction can be made between internal and external validity. Internal validity refers to controlling factors that may cause bias in the results. External (also called ecological) validity is the extent to which results obtained in a laboratory settings correspond to results that would be obtained if the same experiment was carried out in a real-world setting. In the context of sensory analysis, Scriven, (2005) highlights that it is possible to strike a balance between internal and external validity by choice of environment and assessor. A diagram representing this can be seen in Figure 2.4. It is seen that high internal validity is achieved at the cost of low external validity by using expert assessors in a laboratory environment. On the other hand, high external validity is achieved at the cost of low internal validity by using naïve consumers in a home environment.

Scriven, (2005) also discusses the difference between affective and descriptive listening with regards to naïve and expert listeners. It is stated that there are two responses

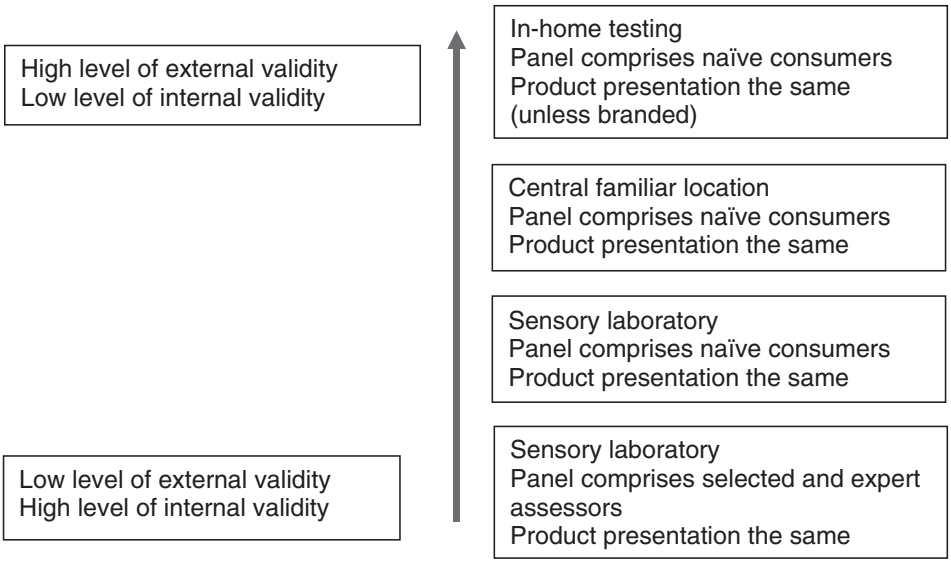


Figure 2.4: The relationship between validity and type of test (Scriven, 2005).

when evaluating a stimuli; the first being to recognise and measure the stimulus and the second being to form a judgement about what has been perceived. It is suggested that naïve listeners are usually only aware of the second response and talk in terms of affective measures. On the other hand, expert listeners are more adept at responding in the first way and can measure constituent attributes of the stimuli in a more objective manner.

2.4.3 *Assessor Categorisation*

As suggested in the discussion on validity, in the field of sensory evaluation there is often a need to categorise assessors by factors such as their acuity, ability and their previous exposure to and knowledge of the type of stimuli in question. However, the terminology used to describe differing levels of skill, expertise and suitability for performing perceptual evaluations in the field of audio is often ambiguous and spread across many standards (Bech and Zacharov, 2006, p. 107). A structure of terms to categorise assessors which is widely employed in the sensory evaluation of food is presented in ISO standard 8586:2012 (ISO, 2012). These terms are related to those often used in the evaluation of audio and, indeed, the structure presented in (ISO, 2012) is reproduced in the audio specific standard ITU-R BS.2300-0 (2014), see Table 2.2. At one end of the scale are naïve assessors (also referred to as untrained or non-expert) who could be considered as the general consumer; they have no previous

Table 2.2: Assessor categorisation terminology based upon ISO 8586 (ISO, 2012), reproduced from ITU-R BS.2300-0 (2014).

Assessor category	Performance description
Assessor	Any person taking part in a sensory test
Naïve assessor	A person who does not meet any particular criterion
Initiated assessor	A person who has already participated in a sensory test
Experienced assessor	Assessor chosen for his/her ability to carry out a sensory test
Expert assessor	Selected assessor with a high degree of sensory sensitivity and experience in sensory methodology, who is able to make consistent and repeatable sensory assessments of various products

experience of sensory evaluation of the type of stimuli in question (e.g. audio). At the other end of the scale are expert assessors (also referred to as trained assessors) who have experience of sensory evaluation and can make consistent and repeatable sensory assessments. It is possible for naïve assessors to develop into expert assessors through training. The effect of this is shown by Bech, (1992) who shows that training reduces the error variance of subjects' ratings of repeated stimuli. An advantage of using expert assessors as reliable quality meters is that they can identify small differences between stimuli and, due to their reliability, usable results can be gathered from relatively few experimental iterations. However, assessor training could lead to important differences between subjects being trained out resulting in subjects providing the answers they are trained to provide as opposed to the answers they would provide otherwise (Berg and Rumsey, 1999). A limitation of using trained listeners is that this may limit the external validity of the results, i.e. it would not necessarily make sense to draw conclusions about the general population based on results from trained listeners as they are not a representative sample.

A range of studies have investigated how evaluations given by naïve and expert listeners compare, several examples of which are described here. Rumsey et al., (2005b) investigated how evaluations given by naïve and experienced listeners differ when assessing band-limiting and down-mixing of multichannel audio. It was found that preference ratings by naïve listeners showed a good correlation to ratings of basic audio quality as given by experienced listeners. This shows it is possible to roughly predict preferences of naïve listeners from ratings given by experienced listeners, which is beneficial as tests with naïve listeners are often much more time consuming. A more

accurate prediction was made when using ratings of timbral fidelity, frontal spatial fidelity and surround spatial fidelity. Whereas there was a good correlation between preference and basic audio quality, it was found that naïve and experienced listeners based their decisions on different aspects of the sound. Experienced listeners gave a higher importance to frontal spatial fidelity than naïve listeners and experienced listeners gave less importance to surround spatial fidelity than naïve listeners.

A study by Guastavino, (2003) (described in (Guastavino and Katz, 2004)) investigated how groups of sound engineers, acousticians and non-experts rate different reproduction methods for reproducing indoor and outdoor sound scenes. A sound field microphone, binaural microphones on a dummy head and a setup of five non-coincident microphones were used and the listeners were asked which recording sounded most like their everyday experiences. It was found that audio engineers gave greater attention to the localisation and precision of the sources, whereas the other two groups based their selection on presence and spatial distribution of sound.

Schinkel-Bielefeld, Lotze, and Nagel, (2013) investigated the difference between experienced and inexperienced listeners in the case of ITU-R BS.1534 (MUSHRA) listening tests for evaluating single-channel speech and audio codecs. It was found that on average the absolute stimuli ratings given by inexperienced listeners were higher than those given by experienced listeners, although the relative ratings were roughly consistent between the two groups. As in Bech, (1992) it was found that inexperienced listeners are less reliable than experienced listeners, meaning that a greater number of inexperienced subjects are needed to gain a comparable confidence interval in comparison to experienced listeners. Additionally, the subjects' listening strategies were studied and it was seen that inexperienced and experienced listeners have different strategies; generally, experienced listeners set more loops of the stimuli and compared more between different stimuli. This was investigated and discussed further in (Schinkel-Bielefeld, 2017).

2.5 GLOBAL JUDGMENT METHODS

Several well known methods for assessing perceived sound quality take the approach of rating all aspects of sound quality in a single judgement. Examples of these are briefly discussed in this section.

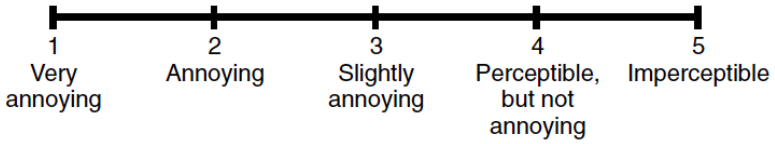


Figure 2.5: ITU-R five-grade continuous impairment scale used in ITU-R BS.1116 (2015b), reproduced from (Bech and Zacharov, 2006).

2.5.1 *ITU-R BS.*1116

ITU-R BS.1116 is the International Telecommunication Union (ITU) standard "methods for the subjective assessment of small impairments in audio systems" (ITU-R, 2015b). As the name suggests, this method is suitable for evaluating systems that introduce small quality impairments (e.g. codecs) and, as such, the standard states that only expert listeners "who have expertise in detecting these small impairments" should be used. To ensure expertise in assessors, pre-screening procedures are suggested (e.g. audiometry) as well as post-screening procedures (e.g. evaluating inconsistencies). The method itself is a "double-blind triple-stimulus with hidden reference" method as it is said that such a method is especially sensitive, stable and permits accurate detection of small impairments. In this method the listener is presented with three stimuli labelled as "A", "B" and "C". Stimulus "A" is always the known reference and the hidden reference and the object are randomly assigned to "B" and "C". The participants' task is to assess the impairments on stimulus "B" compared to "A", and "C" compared to "A" according to the continuous five-grade impairment scale, as shown in Figure 2.5. The five anchors on this scale range from "very annoying" at the low end of the scale (grade 1) to "imperceptible" at the high end of the scale (grade 5). This assessment can take place on a range of attributes, but the standard recommends that "basic audio quality" (BAQ) is evaluated in each case. This is defined as "this single, global attribute is used to judge any and all detected differences between the reference and the object". Other attributes that could be rated include "front image quality" and "impression of surround quality" (in the example of a multichannel system).

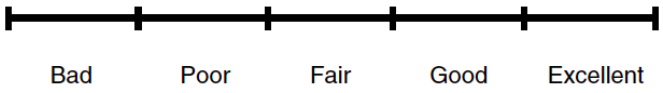
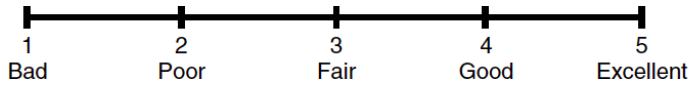


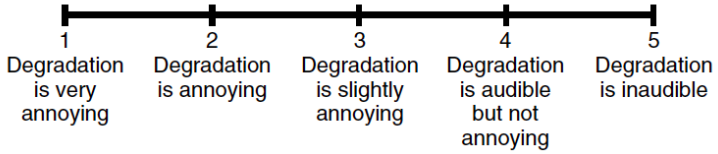
Figure 2.6: ITU-R continuous quality scale (CQS) used in ITU-R BS.1534 (2015c), reproduced from (Bech and Zacharov, 2006).

2.5.2 *ITU-R BS.1534 (MUSHRA)*

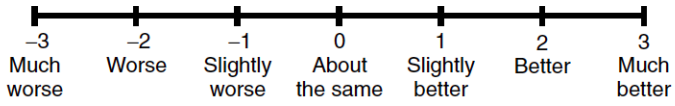
ITU-R BS.1534, better known as MUSHRA (Multi-Stimulus Test with Hidden Reference and Anchor), is the ITU standard "method for the subjective assessment of intermediate quality level of audio systems" (ITU-R, 2015c). Whereas the aforementioned ITU-R BS.1116 is suitable for the assessment of small impairments, it is less suitable for the assessment of lower quality systems as the method is poor at discriminating between small differences in quality at the bottom of the scale. ITU-R BS.1534 is however designed to give reliable and repeatable assessments of systems having audio quality which would normally fall in the lower half of the ITU-R BS.1116 impairment scale. Despite the method not being intended for assessing small impairments, experienced listeners who have "experience in listening to sound in a critical way" are still recommended and pre- and post- screening procedures are outlined. In terms of the method itself, MUSHRA is a "double-blind multi-stimulus test method with hidden reference and hidden anchors". As the method is multi-stimulus, participants are presented with multiple stimuli per trial in order to compare and rate (no more than 12 per trial are recommended). Of these, one is a known reference to which ratings are made, one is a hidden reference, one is a hidden low anchor (a low-pass filtered item with a cut-off frequency of 3.5 kHz), one is a hidden mid anchor (a low-pass filtered item with a cut-off frequency of 7 kHz) and the remaining stimuli are the systems under test (up to 9 per trial). The grading scale on which ratings are made is a continuous quality scale (CQS), which ranges from values of 0 to 100 and is divided into five equal intervals with descriptors ranging from "bad" to "excellent", as shown in Figure 2.6. As with ITU-R BS.1116, it is recommended that the attribute "basic audio quality" is used in each case to judge any and all detected differences between the reference and the object. Again, other attributes that could be rated include "front image quality" and "impression of surround quality".



(a) Absolute category rating (ACR).



(b) Degradation category rating (DCR).



(c) Comparison category rating (CCR).

Figure 2.7: Rating scales used in ITU-T P.800 (1996), reproduced from (Bech and Zacharov, 2006).

2.5.3 *ITU-T P.800*

ITU-T P.800 is the ITU standard "methods for subjective determination of transmission quality" (ITU-T, 1996). The scope of this standard is to provide approved methods suitable for determining how satisfactorily given telephone connections may be expected to perform. The methods are intended to be used with any form of degradation including transmission errors, talker echo, distortion and environmental noise, among others. Unlike the two standards discussed above, the methods outlined in ITU-T P.800 are intended for use with naïve listeners who have not been directly involved in work related to the assessment of relevant systems and who have not participated in any subjective test for at least the previous six months. Three methods are outlined: the "absolute category rating method" (ACR), the "degradation category rating" (DCR) method and the "comparison category rating" (CCR) method.

The ACR method is a simple single stimulus method whereby participants rate stimuli on a five-point category-judgement scale. The most commonly used scale is a mean opinion score (MOS) scale on which participants are asked to rate "listening-quality opinion". Such a scale ranges from "bad" to "excellent", as seen in Figure 2.7a. Other possible scales include a listening-effort scale and a loudness-preference scale.

The DCR method involves presenting pairs of stimuli to participants; as such it has a higher sensitivity than the single stimulus ACR method. For each pair of stimuli, e.g. "A" and ''B", the quality reference sample ("A") is presented first followed by the same sample processed by the system under evaluation ("B"). Additionally, "null-pairs" such as "A"-"A" are included to check for consistency. A five-point category-judgement scale is used to rate the level of degradation introduced by the system under study in comparison to the reference. Possible ratings range from "degradation is very annoying" to "degradation is inaudible", as shown in Figure 2.7b.

Whereas with the DCR method where ratings are always made of a quality degradation of a system in comparison to the reference, the CCR method allows for the rating of either a quality degradation or a quality improvement in comparison to the reference. The CCR procedure is similar to that of the DCR method, the key main difference being that with the CCR procedure the order of the processed and unprocessed samples is chosen at random for each trial. Listeners are tasked with comparing the quality of the second stimuli to that of the first on a scale ranging from "much worse" to "much better", see Figure 2.7c. It is noted that with such a scale, listeners are providing judgments of both which sample has better quality and also by how much.

2.5.4 *Affective Measures*

The previously discussed global judgment methods could generally be regarded as using perceptual measures with regards to the filter model presented in Figure 2.3. That is, the responses sought do not take into account cognitive factors such as mood, context, emotion, background and expectation. Global measures that do take into account such factors can be referred to as affective (or hedonic) measures. Such measures are often used in the field of food evaluation (Lawless and Heymann, 2010), although they are also applicable to the evaluation of audio. Consumers (or naïve listeners in the realm of audio evaluation) are generally used for affective testing and, due to the high variability of individual preferences, an increased sample size should be used in comparison to non-affective global judgment tests. Forms of affective testing include preference tests (by paired or multiple comparisons), acceptance tests (for example by means of the scale presented in Figure 2.8) and appropriateness tests.

An affective measure specific to the field of audio evaluation is "overall listening experience" (OLE). This term, recently used and defined by Schoeffler et al. over a range

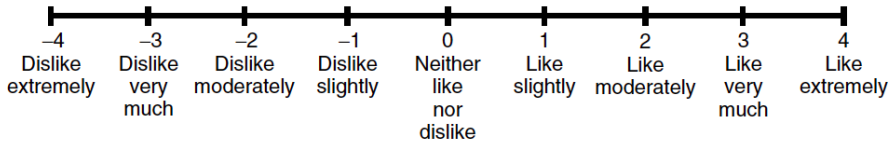


Figure 2.8: Nine-point hedonic scale (Peryam and Girardot, 1952), reproduced from (Bech and Zacharov, 2006).

of studies, is described as being the quality of experience in the context of audio consumption and is intended to include all possible factors that may influence listeners' ratings of stimuli (Schoeffler, Silzle, and Herre, 2017).[1] Possible influencing factors could include the song, lyrics, audio quality, the listener's mood and the reproduction system. OLE has been used in a range of studies including investigations on the influence of timbral audio quality on OLE (Schoeffler, Edler, and Herre, 2013; Schoeffler and Herre, 2013), the influence of up-/down-mixes on OLE (Schoeffler, Adami, and Herre, 2014), the influence of single-/multi-channel systems on OLE (Schoeffler, Conrad, and Herre, 2014) and for the evaluation of 3D audio systems (Schoeffler, Silzle, and Herre, 2017). Furthermore, comparisons between OLE and basic audio quality have been made (Schoeffler and Herre, 2016) in which it was seen that OLE can produce comparable results to basic audio quality. To assess OLE, participants are asked to rate stimuli on a five-star Likert scale taking everything into consideration that is important to them (e.g. quality, content etc.). Ratings are first given for reference conditions (i.e. unprocessed stimuli) and these act as a measure of how much participants like each song without taking any processing into account. These ratings are known as "basic item ratings". Secondly, the conditions to be tested (e.g. different reproduction methods) are rated and these are known as "item ratings". It is then possible to compare the basic item ratings with the item ratings so as to evaluate how much the different conditions influence the overall listening experience.

2.6 ATTRIBUTE-BASED METHODS

An alternative approach to sound quality assessment is to rate stimuli with respect to relevant attributes. The aim of such an approach is to define a collection of attribute scales that can sufficiently describe similarities and differences between stimuli. Bech, (1999) defines an auditory attribute as "a perceived characteristic of a sound stimulus,

1 See (Schoeffler, 2017) for a comprehensive discussion on OLE-based studies.

for example pitch and loudness". It is also explained by Bech that it can be assumed that auditory attributes can be combined to form an overall preference judgement through preference mapping.

The background for attribute-based sensory evaluation can be found in the field of food science, hence a lot of techniques used in attribute-based sound evaluation originate from this field. In terms of perceptual attributes of sound it was the area of concert hall acoustic evaluation that produced some of the earliest work. For example, Sabine, (1900) was interested in what makes good listening conditions in auditoria and identified three contributing factors; (i) loudness, (ii) distortion of complex sounds: interference and resonance, and (iii) confusion: reverberation, echo and extraneous sounds. Perhaps the earliest work on multichannel audio involving perceptual attributes is that of Nakayama et al., (1971). In this study, subjective effects of one to eight-channel reproductions were investigated by recording two popular music extracts at a concert hall and reproducing them over various loudspeaker setups in an anechoic chamber. Preference judgements of each reproduction and similarity judgements among them were made by listeners and multidimensional analysis was applied to the data. It was found that the multichannel recording and reproduction of music from such an acoustical setting is characterised by three sensory features: fulness, clearness and depth of the image sources. Other early work on perceptual attributes of sound reproduction include that by Eisler, (1966), Staffeldt, (1974), Gabrielsson, (1979) and Toole, (1985).

An important early study in which several key points are made on the topic of attributes and the assessment of sound quality in general is that of Letowski, (1989). In this study a distinction is made between the terms "sound character" and "sound quality". It is suggested that sound quality should include preferential and emotive responses whereas sound character should be a purely descriptive measure. The definition of sound quality is given as (Letowski, 1989, p. 6):

> "Sound quality is that assessment of auditory image in terms of which the listener can express satisfaction or dissatisfaction with that image. Sound quality can be judged by comparing images produced by several external stimuli or by referencing a perceived image to the concept residing in the listener's memory."

Furthermore it is suggested that sound quality is a multidimensional entity and that to sufficiently describe an auditory image a multidimensional assessment is needed. With this in mind, Letowski introduced a hierarchical system of auditory sensations

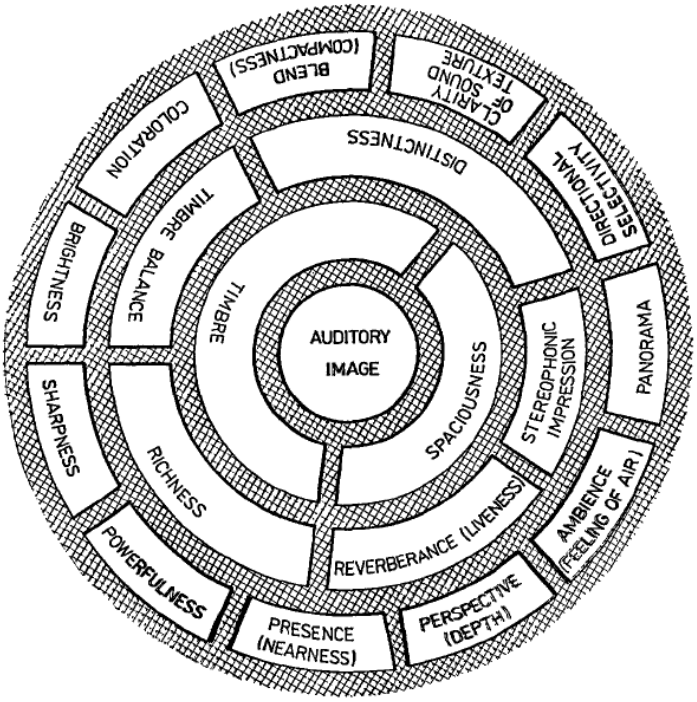


Figure 2.9: Letwoski's MURAL (Letowski, 1989).

called MURAL, which stands for MUltilevel auditoRy Assessment Language, Figure 2.9. This was not intended to be a final and complete model, although some key ideas can still be seen, namely the differentiation between timbral and spatial attributes of an auditory image. Letowski defines timbre as "that attribute of auditory image in terms of which the listener judges the spectral character of sound" and spaciousness as "that attribute of auditory image in terms of which the listener judges the distribution of sound sources and the size of acoustical space".

A diagram outlining how various subsets and related attributes relate to total perceived quality is presented by Berg and Rumsey, (2003) and is shown here in Figure 2.10. It suggests that total audio quality consists of the subsets "timbral quality", "spatial quality" and "technical quality" and these are further split into "timbral attributes", "spatial attributes" and "technical attributes". Clear similarities between this and Letowski's MURAL can be seen, although this schematic has the added subset of "technical quality", which includes factors such as distortion, hiss and hum.

The relative importance of timbral quality and spatial quality on total perceived quality was investigated by Rumsey et al., (2005a). In this study multichannel audio excerpts were degraded both with respect to timbre (bandwidth limitation) and spatial information (down-mixing) and then evaluated using the attributes of basic audio

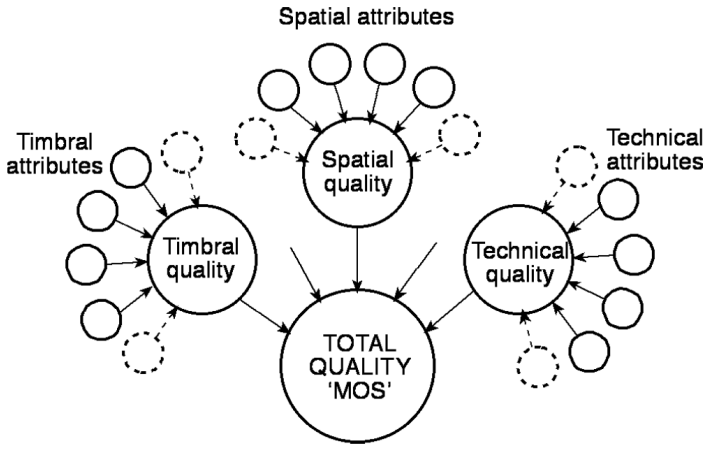


Figure 2.10: Relations between total audio quality and its subsets and attributes (Berg and Rumsey, 2003).

quality, timbral fidelity, frontal spatial fidelity, and surround spatial fidelity. It was found that whilst timbral fidelity was the prominent influencing factor on general audio quality, spatial audio quality constituted approximately 30% of the overall quality rating. These tests were only undertaken by experienced listeners so may not be generalisable. However, the results still show that spatial quality is an important factor worthy of attention.

2.6.1 *Considerations on Attribute Selection*

As previously mentioned, it is possible and often desirable to evaluate sound with respect to various attributes. Such attributes are generally expressed through verbal descriptors for use in evaluation scales. A key requirement of these descriptors is that they should reflect the perceived audible sensations as closely as possible. This therefore makes the task of identifying which attributes to assess an important one. According to Berg, (2006), as well as reflecting the perceived audible sensations of the listener, the attributes should "be clear and unambiguous to allow for a common understanding across subjects and scales should differentiate between stimuli". Additionally, Berg goes on to say "if no overlap of the scales is desired, the scales should also be orthogonal". This is also briefly discussed by Rumsey, (1998) who says that it is open to discussion whether attributes need to be orthogonal. He identifies that while it is mathematically neat for the dimensionality of space perception to be reduced to as few dimensions as possible, it is also important that the scales or dimensions defined are meaningful.

Several categorisations in terms of attribute generation are made by Berg and Rumsey, (1999). A distinction is made between "provided constructs", which are terms or definitions provided by the experimenter, and "elicited constructs", which are terms or definitions suggested by or elicited from the subject. Additionally, it is suggested that the various methods for eliciting attribute scales in subjective tests can be split into three groups: those with the intention of arriving at a common set of attributes for grading by all subjects (consensus vocabulary (CV) methods), those that are based on free categorisation or individualised scales (individual vocabulary (IV) methods) and those which use multidimensional analysis based on non-semantic similarity/difference relationships between stimuli. An advantage of using a common set of of attributes is said to be that the results from multiple subjects can be statistically analysed and inferences can be drawn regarding the preferences of the general population. Individualised scales have the advantage of lack of bias and a greater opportunity for personal reflection, without the need for subject training. Finally, the third group has the advantage of a lack of bias but is said to have problems in terms of interpretation and application in practice. The differences between CV and IV techniques are discussed in greater detail by Pedersen and Zacharov, (2015).

2.6.2 *Provided Construct Methods*

Provided constructs, that is constructs that are provided and defined by the examiner, are sometimes used when previous experiments of a similar context have shown them to be useful. It is possible that the provided constructs in a given experiment were derived from elicitation in previous experiments. An advantage of using provided constructs is that, when using a well-defined scale, the experimenter can focus the subject's attention on certain sensations or attributes of which the experimenter wants to study. Moreover, provided construct methods can be more time-efficient than elicited construct methods if no training on the attributes is needed. A limitation of using provided constructs is that it is possible that the subjects may not be able to relate to the given attributes (Berg, 2006). Limitations of using provided constructs are also touched upon by Berg and Rumsey, (1999). A quote from Kjeldsen, (1998) is used to point out a limitation of semantic differential method based on provided constructs - "an obvious limitation of this type of measures is, that you only get an answer to what you ask".

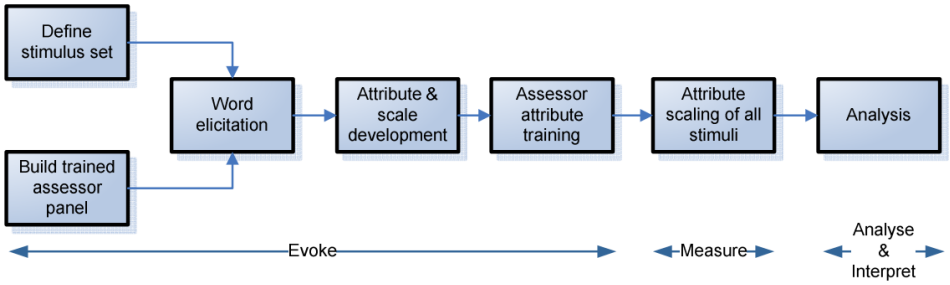


Figure 2.11: Overview of the descriptive analysis process (Pedersen and Zacharov, 2008).

2.6.3 *Elicited Construct Methods*

A range of methods exist for the elicitation of constructs and, as mentioned, these can be grouped into consensus and individual vocabulary methods. An overview of some of these methods is given here.

Quantitative Descriptive Analysis (QDA), developed by Stone et al., (1974), is an elicitation method that results in a common scale of attributes. The first step of this method involves the selection of an expert assessor panel based on their discriminatory ability. Under the guidance of a panel leader, these assessors then generate a descriptive language which is subsequently used in the grading of stimuli. An overview of the process is presented in Figure 2.11. It should be noted that this process involves an "assessor attribute training" section. This is due to the fact that this method is of an objective manner whereby the subjects are used as reliable quality meters. To improve the subjects' reliability they need to be calibrated and this is the purpose of the assessor training. QDA could be considered as possessing features of both elicited construct techniques and provided construct techniques. Each assessor has the opportunity to influence the attribute list through their personal attributes and definitions, although assessors are also influenced and biased by each others given attributes and definitions. Examples of studies that have used QDA include (Koivuniemi and Zacharov, 2001), (Martin and Bech, 2005) and (Lorho, 2005a).

Free-Choice Profiling (FCP) is a method that results in a set of individual attributes for each subject. Its advantages are that panel training is not necessary meaning that consumers can be directly used for the listening tests, there is no discussion between subjects and there is no experimenter influence biasing the results. For meaningful results to be obtained from FCP advanced statistical methods need to be employed. The method was developed in the 1980's by sensory scientists, see (Williams and

Langron, 1984), and to analyse the results a statistical technique called Procrustes analysis is used, see (Dijksterhuis, 1996). The method works by allowing subjects to use their own words to describe and evaluate the stimuli and through the statistical analysis it is possible for the examiner to group terms that appear to relate to the same sensation. Examples of studies inspired by FCP include (Lorho, 2005b) (also inspired by the Flash Profile method) and (Guastavino and Katz, 2004). In the study by Lorho, (2005b) it was found that consensus methods took between 20 and 60 hours for vocabulary development whereas the individual method used took only three hours.

Another approach to elicit individual attribute scales is the Repertory Grid Technique (RGT). This technique was originally developed by Kelly, (1955) and was subsequently brought to the field of spatial audio evaluation by Berg and Rumsey, (1999). The idea behind RGT is for subjects to develop a personal set of constructs on which to rate the stimuli as, according to Berg and Rumsey, (1999), subjects have been shown to be more reliable when using their own language than that of others. The technique could be said to be split into three main steps: elicitation, scaling and data analysis.

- *Elicitation*: The elicitation phase of RGT is generally based on a triadic structure; the subject is presented with three randomly selected stimuli and is asked in which way two of them are alike and different from the third. This results in a bipolar construct which describes the similarity and difference between the chosen grouping. This is repeated until all combinations of stimuli have been presented.

- *Scaling*: The bipolar constructs, as elicited in the first stage, are then used as end points of multiple scales on which to rate the sound stimuli. Each subject uses their bipolar scales to rate the stimuli and this results in a matrix grid for each subject with the scales and score for each stimuli.

- *Data Analysis*: Berg and Rumsey, (1999) suggest various methods for the analysis of the matrix grid. These include cluster analysis, principal component analysis (PCA) and and rank order correlation.

A disadvantage of using a triadic structure for rating stimuli is that differences between two stimuli may be overlooked if they are presented with a third stimuli which is even more dissimilar. To overcome this limitation Choisel and Wickelmaier, (2006) investigated the use of a pairwise structure whereby pairs of stimuli are compared. It was found that both methods produced a comparable number of descriptors, but

it was noted that it was easier to interpret the bipolar constructs as end points of rating scales in the case of pairwise comparison. Another disadvantage of RGT is the possibility of experimenter bias when interpreting the results. To limit this, the subjects could be consulted when the interpretation takes place. Examples of other studies that have utilised RGT include (Berg and Rumsey, 2000), (Berg and Rumsey, 2002) and (Geier et al., 2010). In order to make RGT more employable, Berg, (2005) developed a software tool called OPAQUE with the purpose of aiding the elicitation process, the scaling of stimuli and the analysis and presentation of the results.

The elicitation methods presented so far can all be classed as direct elicitation methods. With such methods it is assumed that subjects can represent perceived sensations through the use of a verbal descriptors. It is possible that subjects may perceive a sensation which is not suited to being described verbally and this presents a possible limitation of direct elicitation methods. Indirect elicitation methods on the other hand aim to uncover salient perceptual attributes without the subject directly identifying them. Multidimensional Scaling (MDS) is such a method and is based on difference and similarity ratings between stimuli. The technique involves asking subjects to rate stimuli on a scale of dissimilarity, to then interpret these ratings using a multidimensional plot and subsequently extract various dimensions present in the stimuli. In terms of spatial audio MDS is discussed by Choisel and Wickelmaier, (2006) and used by Martens and Zacharov, (2000). Another indirect elicitation method is Perceptual Structure Analysis (PSA), introduced by Choisel and Wickelmaier, (2006). Similarly to RGT, in PSA the subjects are presented the stimuli in a triadic format. After each triad the subjects are then asked "do sounds 'A' and 'B' share a common feature that sound 'C' does not have?" and the subjects reply with a simple "yes" or "no". By presenting all combinations of stimuli to the subjects it is then possible for the experimenter to extract the auditory features underlying the responses. Projective Mapping is yet another indirect elicitation method. Originally introduced for the sensory evaluation of food, it has been applied to the evaluation of loudspeakers by Giacalone et al., (2017) and involves positioning stimuli labels on a sheet of paper (or GUI) in such a way that two stimuli should be placed close if they are perceived as similar and far apart if they are perceived as different. Analysis is by means of Multiple Factor Analysis and results in a perceptual map of the stimuli under test.

Attribute	Description
Spatial fidelity	Degree with which spatial attributes agree with reference
Spaciousness	Perceived size of environment
Width	Individual or apparent source width
Ensemble Width	Width of the set of sources present in the scene
Envelopment	Degree to which the auditory scene is enveloping the listener
Depth	Sense of perspective in the auditory scene as a whole
Distance	Distance between listener and auditory event
Externalisation	Degree to which the auditory event is localised in- or outside the head
Localisation	Measure of how well a spatial location can be attributed to an auditory event
Robustness	Degree to which the position of an auditory event changes with listener movements
Stability	Degree to which the location of an auditory event changes over time

(a) Spatial attributes

Attribute	Description
Timbral fidelity	Degree to which timbral attributes agree with reference
Colouration	Timbre-change considered as degradation of auditory event
Timbre, colour of tone	Timbre of the auditory event(s)
Volume, richness	Perceived "thickness"
Brightness	Perceived brightness or darkness (dullness)
Clarity	Absence of distortion, clean sound
Distortion, artefacts	Noise or other disturbances in auditory event

(b) Timbral attributes

Table 2.3: Examples of (a) spatial and (b) timbral attributes used to describe auditory events in the context of sound reproduction systems (Spors et al., 2013).

2.6.4 *An Overview of Perceptual Attributes*

Due to the nature of elicitation of attributes it is not surprising that different experiments result in different elicited attributes. Factors affecting the attributes that are elicited could include the reproduction system under evaluation, the stimuli used and the background of the subjects. Despite the fact that there is no standard set of attributes for perceptual spatial audio evaluation, many of the attributes identified in various experiments are often very similar. Spors et al., (2013) reviewed the literature for common spatial and timbral attributes found in perceptual experiments and these attributes along with their respective meanings are presented in Table 2.3. It should be noted that these meanings are not as unambiguously defined as one may imag-

ine, with different studies often using different definitions. For example, Berg, (2009) identifies conflicting and contrasting definitions of envelopment and concludes that the terms "envelopment" and "immersion" are both used inconsistently throughout the literature. In terms of the broader context of knowledge elicitation, this possible ambiguity in terminology is highlighted by Shaw and Gaines, (1989, p. 341):

> "One problem of eliciting knowledge from several experts is that experts may share only parts of their terminologies and conceptual systems. Experts may use the same term for different concepts, use different terms for the same concept, use the same term for the same concept, or use different terms and have different concepts."

An extensive list of 48 attributes was developed by Lindau et al., (2014) in the context of virtual auditory environments. The attributes were generated by a focus group of 20 experts for virtual acoustics and involved 56 hours of discussions spread over six months. It is noted that the discussion time for this method was similar to QDA and RGT elicitation methods. The 48 attributes were split into eight categories: timbre, tonalness, geometry, room, time behaviour, dynamics, artefacts and general impression.

A comparison of elicited attributes from various experiments covering different contexts and using different elicitation techniques was made by Pedersen and Zacharov, (2008). Attributes were included from studies that evaluated multichannel reproduction as well as headphone reproduction and the different elicitation methods included RGT, QDA, descriptive analysis and PSA. It was noted that as many attributes were similar despite different elicitation methods being used, it is possible that there may be potential to define a core set of attributes.

The sound wheel for reproduced sound (Pedersen and Zacharov, 2008), see Figure 2.12, aims to provide such a common terminology (or "lexicon") for the characterisation of sound quality in loudspeakers, headphones and other sound reproduction systems. The wheel format, which has been used in other areas of sensory evaluation such as the wine industry, structures sensory characteristics hierarchically; towards the centre of the wheel terms describing groups of similar sensory attributes are found (e.g. "timbre") and on the edge of the wheel more specific attributes are found (e.g. "boomy"). It is mentioned that a complete lexicon should cover all relevant attributes for the domain in question, although only a subset of attributes will typically be used in a specific test with a limited number of products. Attributes for the sound wheel presented above were elicited using mono and stereo systems only. It could therefore

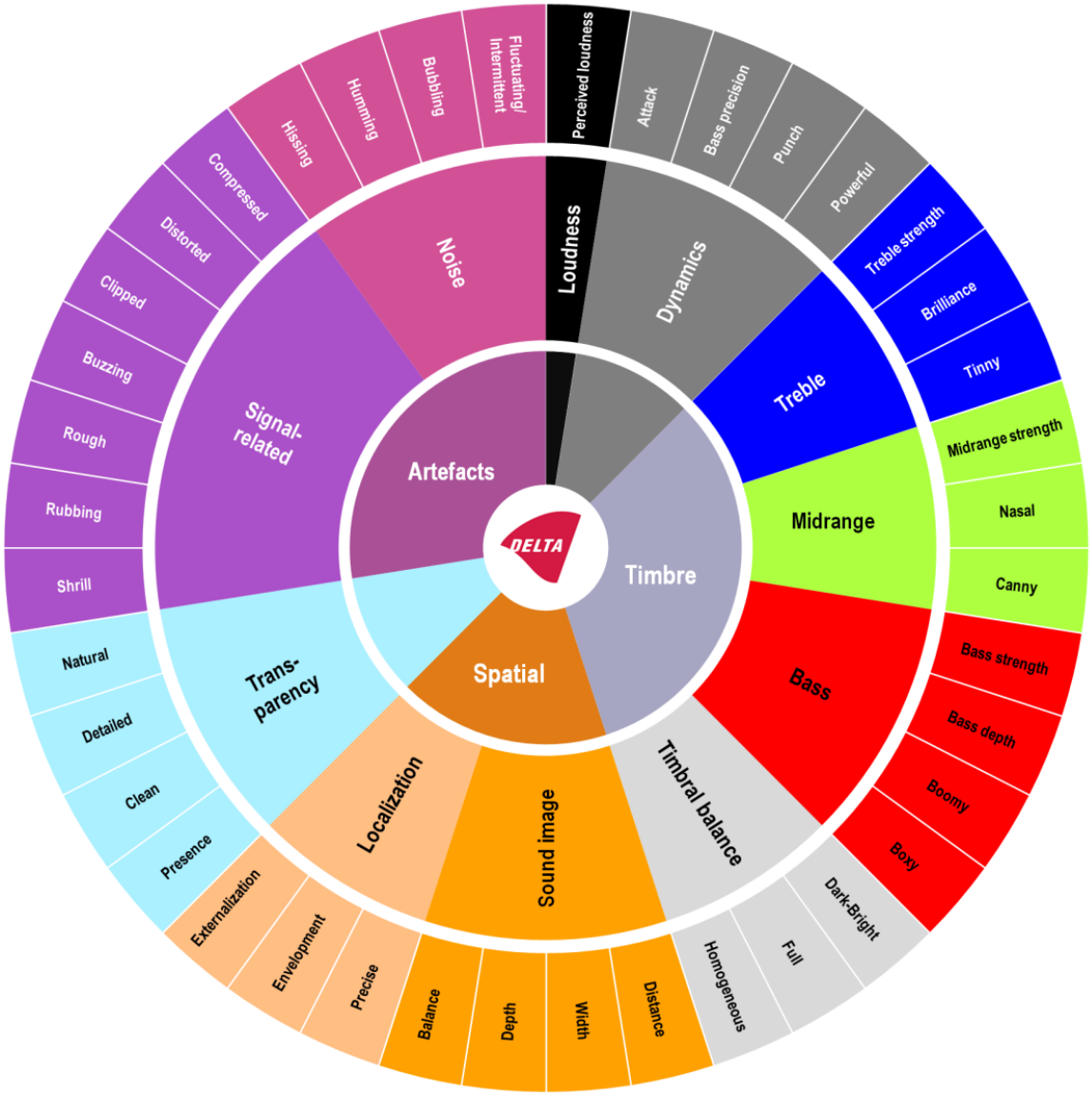


Figure 2.12: Sound wheel for reproduced sound (Pedersen and Zacharov, 2015).

be expected that when including multichannel and advanced sound systems in the elicitation process a higher number of spatial attributes would be included.

2.7 COMBINING GLOBAL AND ATTRIBUTE-BASED METHODS

In the previous sections a range of methods that use global judgments and a range of methods that use attribute-based judgments to evaluate audio quality have been outlined. A variety of studies exist that combine these types of evaluation with the intention of gaining a greater understanding of the stimuli under study. A key feature often found in such studies is preference mapping. The aim of preference mapping is to relate perceptual attributes to preference ratings so that knowledge about which attributes contribute most to consumer ratings is gained.

Zacharov and Koivuniemi, (2001) introduced a method called Audio Descriptive Analysis and Mapping (ADAM) in the context of perceptual evaluation of spatial audio systems. In this method a preference rating task is completed by naïve participants in a paired comparison format, a language development task is performed with trained participants, a discussion phase creates a common descriptive language which is then used in an attribute rating stage by trained participants. Finally, partial least squares regression is used to map the subjective preference ratings to the attributes.

Choisel and Wickelmaier, (2007) conducted an experiment with naïve participants with the aim of quantifying the auditory attributes that underlie listener preference for reproduced multichannel sound. They collected preference ratings via paired comparison judgments and utilised attributes elicited in a previous study to develop ratio scales from probabilistic choice models. Principal components derived from the quantified attributes were then used to predict overall preference.

Zacharov et al., (2016) presented a method for the assessment of next generation audio systems called the Multiple Stimulus Ideal Profile Method (MS-IPM). Originally developed in the perfume industry and later applied to hearing aid applications, the method aims to relate overall quality, attribute ratings, and also an "ideal profile". This ideal profile is obtained by asking participants to give an ideal level of each attribute on which the stimuli are being assessed. With regards to attribute elicitation, in this example of the method four specialised expert assessors selected six attributes on which the stimuli were to be rated by the experienced participants.

Francombe and colleagues recently presented a series of papers (Francombe et al., 2016; Francombe, Brookes, and Mason, 2017; Francombe et al., 2017b) on a method

to evaluate spatial audio reproduction systems by combing preference ratings with attribute data. The method involves a paired comparison procedure to gather preference ratings alongside a free elicitation task to elicit perceptual differences between stimuli. Due to the simple paired comparison format the method is suitable for both experienced and inexperienced listeners. Automatic text clustering was used to reduce redundancy in the attribute data and the elicited attributes were further refined by means of group discussions. To analyse the importance of the various attributes on preference ratings a metric called "attribute score" was developed, which quantifies the importance of each attribute by considering the frequency with which it was used as well as the size of the preference judgments alongside which it was used.

Studies also exist that relate preference ratings with sensory profiling for perceptual evaluation in other fields of acoustic research. For example, Mattila, (2001) combined descriptive analysis in the form of paired comparison attribute elicitation and panel discussions, with overall quality judgements for the evaluation of speech quality in mobile communications. In the context of concert hall acoustics, Lokki et al., (2012) conducted an individual vocabulary profiling procedure with a triad-based elicitation stage and single stimulus attribute rating stage and combined these attribute ratings with preference ratings via preference mapping.

2.8 SOUND QUALITY MODELS

Listening tests are often very resource intensive. It is therefore desirable to generate a model which can predict sound quality without the need for listening tests. One technique to do this is to calculate quality based on a comparison between an unprocessed reference and a processed signal. Such models are known as full-reference models and examples include PESQ (ITU-T, 2001), POLQA (ITU-T, 2011) and PEAQ (ITU-R, 2001). PESQ (Perceived Evaluation Speech Quality) is used for speech assessment, as is its successor POLQA (Perceptual Objective Listening Quality Assessment). PEAQ (Perceptual Evaluation of Audio Quality) on the other hand is aimed at evaluating audio quality. These are perceptual psycho-acoustic models that aim to act like an artificial ear by using models of the auditory periphery. In this way, effects such as masking and cognitive effects can be taken into account. Such models are restricted to a limited set of conditions (Raake and Blauert, 2013) and are therefore not suitable for spatial sound assessment.

A model designed with the context of spatial sound quality in mind is QESTRAL (Conetta et al., 2008; Dewhirst et al., 2008; Jackson et al., 2008; Rumsey et al., 2008). It is based upon a scene-based evaluation after Rumsey's scene-based paradigm (Rumsey, 2002) and takes into account spatial distortions such as source location, width and envelopment as well as others. It also includes foreground-background separation as in the scene-based paradigm. Like the previous models mentioned it uses a reference and was calibrated by a large database of listening test results.

Raake and Blauert, (2013) point out that these models do not take into consideration the "world knowledge of listeners" and thus individual factors, such as mood, expectation, and experience are not accounted for. They also refer to Thiede et al., (2000) to make the point that without knowing the "ideal audio signal... in the mind of the listener" it is very difficult to model listener behaviour.

2.9 EVALUATION OF NEXT GENERATION AUDIO: THE STATE OF THE ART

The sections presented above have been predominantly concerned with techniques to evaluate the quality of audio technology. In this section, we turn our focus to look at results of studies specifically concerned with the quality assessment of next generation audio.

2.9.1 *Immersive Audio*

Several recent studies have investigated the perceived quality of immersive loudspeaker setups, a selection of which are discussed here. Kim, Lee, and Pulkki, (2010) compared a 22.2-channel system, which has nine elevated channels, with systems with four, three, two and zero elevated channels using an ITU-R BS.1534 (MUSHRA) method. Using material mixed on a 22.2-channel system, they found that whist the inclusion of elevated loudspeakers increased the perceived quality, a $7+3$ system with three elevated channels was similarly rated to the 22.2-channel reference. This result suggests that a large number of height channels may not be necessary to provide a good experience. These results are however only based off overall quality judgments so may not be applicable to certain attributes.

Silzle et al., (2011) conducted two experiments to compare a 22.2 system with two-, five- and nine- $(5+4)$ channel systems; one experiment without a reference and one with (the 22.2 system). Both experiments revealed that listeners preferred the systems

with height channels, although the extent of this was content dependent. When an explicit reference was used, it was seen that the 22.2 system was rated with significantly higher quality than the nine-channel system. However, without a reference minimal differences between the two systems were seen. These results illustrate the influence of a reference in quality evaluation tasks and indeed it is mentioned by the authors that a paired comparison test could lead to more "unbiased" results.

Unlike the previous studies which used audio-only content, Cobos et al., (2015) assessed the subjective quality of multichannel audio accompanied with video for representative broadcast genres. Stereo, 5.1 and 7.1 surround, 10.1 $(7+3)$ and binaural systems were compared based on an ITU-R BS.1286 method, including both absolute category rating and paired comparison tests. The systems were compared using a range of attributes including frontal sound image quality, impression of surround quality, correlation of source positions derived from visual and audible cues, correlation of spatial impressions between sound and picture, and basic audio quality. Results showed that the only attributes significantly influenced by the reproduction systems were surround and basic audio quality. The with-height 10.1 system was consistently preferred over the horizontal-only systems, the differences highlighted most with the absolute category rating method. The binaural items were rated poorly compared to the other systems with no significant differences seen between these and the stereo items. This was attributed to "the 'inside the head' effect, headphone discomfort and the lack of bass power". As with the other studies, the content had a strong influence on the perceived quality.

Schoeffler, Silzle, and Herre, (2017) compared 22.2, 5.1 surround and stereo systems using both basic audio quality (with reference) and overall listening experience (without reference) methods. As with the previous studies, the with-height 22.2 system was rated as having higher quality than the horizontal-only systems, and as with the study by Silzle et al., (2011), the use of a reference expanded the difference between the 22.2 and other systems. In the case of the OLE method, which did not use a reference, the difference between the 22.2 and 5.1 surround systems were seen to be much smaller.

Most recently, Francombe et al., (2017b) performed a paired comparison alongside a free elicitation task in order to relate listener preference with relevant perceptual attributes for a range of immersive systems. Eight systems were compared ranging from low quality mono to 22-channel surround, with content including music, sport and film genres. In terms of preference for the reproduction methods, both experienced and inexperienced listeners rated surround higher than stereo and stereo higher than

mono. However, there was little difference between the five-channel and nine-channel systems and the 22-channel system was rated lower than the five- and nine-channel systems. This result was somewhat dependent upon content; it was speculated that 22-channel was less preferred as the content was less suitable. These results are in line with results from Silzle et al., (2011) who showed that, without a reference, there were minimal differences in preference between 22-channel and nine-channel systems. With regards to the relevant perceptual attributes, the attributes "amount of distortion", "bandwidth" and "output quality" were important when distinguishing between low and high quality systems, whereas the attributes "enveloping" and "horizontal width" were used very frequently and showed a strong relationship with preference scores for the higher quality systems.

2.9.2 *Object-Based Audio*

Whereas a range of studies have investigated the quality of next generation immersive systems, fewer have evaluated the benefits provided by object-based audio. Perhaps some of the most relevant studies concerning object-based audio are those by Churnside, (2016). The PhD thesis by Churnside describes three case studies, each considering the impact of using object-based audio on the creative process, production workflow and audience experience. The first of these analysed the audience's use of the ability to personalise the mix of a live football match. Results from this study were previously mentioned in Section 2.2.3. The second study included subjective tests investigating preferences for different mixes of foreground versus background audio levels across different genres and loudspeaker layouts. Results from this study showed that there was no clustering of listeners based on their preference of foreground versus background balances and also that there was significant variation of foreground and background balance preference between loudspeaker layouts. The final study analysed the benefits of being able to adapt the story of a drama so that it is set in a location that is familiar to the listener. Results from this study showed that the tailored version increased the audience's enjoyment.

Shirley et al., (2017) investigated the benefits of object-based audio in the context of improving television sound for hearing impaired people. In their experiments, participants (14 out of 19 having some degree of hearing impairment) could personalise audio levels for the four object-categories of speech, music, background effects and foreground effects related to on-screen events. It was seen that there was a large varia-

tion in preference across participants, although for some hearing impaired people the ability to personalise the four object-categories substantially improved the viewing experience.

Whereas these studies illustrate the range of potential benefits offered by object-based audio, there are many potential benefits that have not yet been studied.

2.10 DISCUSSION

This chapter has presented a review of literature related to audio reproduction technology and its evaluation. An overview of the progression of audio technology from mono to next generation was given, in which it was seen that the next generation of audio technology includes advancements such as object-based and immersive reproduction. These advancements in technology also represent a shift in experience towards immersive, interactive and personalised experiences; the various next generation reproduction methods focus on providing immersive experiences and object-based audio allows for adaptive and personalised content. Furthermore, behavioural trends, such as ubiquitous listening and the extensive use of audio devices with significant processing capabilities, are key factors that will likely play a part in the progression of next generation audio. Ubiquitous listening could place an emphasis on immersive headphone reproduction as well as adaptation of content to suit the environment. The use of devices with significant processing capabilities facilitates the adaptation of content on mobile devices as well as facilitating certain immersive reproduction techniques, such as personalised, dynamic binaural reproduction on mobile devices and media device orchestration.

As the experience provided by audio technology evolves, and also as potential listening environments change, the need to reconsider the methods by which such technology should be evaluated arises. A large section of this chapter explored the topic of audio quality evaluation. The most prominent methods of audio evaluation are those outlined by the ITU. These methods typically require experienced listeners in highly controlled environments to rate stimuli with the global measure of basic audio quality, often in comparison to a high quality reference. These methods are evidently very useful for certain applications, such as codec evaluations, yet for the evaluation of innovative technologies that have the potential of providing new experiences to the user (e.g. object-based and immersive audio), a different approach needs to be taken. Reference-based methods are not so suitable for evaluating technology that de-

livers innovative experiences as a high quality reference is often not available. It could also be argued that affective global measures (e.g. preference) are more suitable than perceptual global measures (e.g. basic audio quality) for evaluating certain next generation audio technologies, as it is possible that different technologies may be of equally high "quality" yet provide different experiences, thus leading to different ratings of preference. Furthermore, when evaluating technology that provides new experiences it is desirable to know which attributes influence overall preference the most so that developers of new technology can utilise this to design for a high quality of experience. This could be achieved by combining global and attribute-based methods in order to determine the relationships between listener preference and the attributes that play a role in the formation of preference for the relevant content and contexts.

The possible advantages to the listening experience provided by next generation audio also require the consideration of factors that are often overlooked in current audio quality evaluation studies. Object-based audio allows for content adaptation to suit the environment. Combined with the trend towards ubiquitous listening, this means that considering the context of use and associated contextual factors is necessary when evaluating certain applications of next generation audio. Additionally, object-based audio allows for content adaptation to suit the user and it is therefore necessary to also consider user factors. These considerations further strengthen the stance that the current standardised methods are not well-suited for the evaluation of next generation audio, as standardised audio quality evaluation methods aim to reduce context and user effects by conducting studies in highly controlled environments and with experienced assessors.

As discussed in Chapter 1, quality of experience is a measure of quality associated with the "enjoyment" of a service or system and one that considers context and user effects. Considering the above discussion on the requirements of next generation audio quality evaluation methods, QoE is therefore a concept that is highly relevant. In the following chapter, the concept of QoE and its suitability for next generation audio evaluation is explored in greater detail.

2.11 SUMMARY

The aim of this chapter was to present and discuss the literature related to audio reproduction technology and its evaluation on which this thesis is based. To begin, an overview of the progression of audio technology from mono to next generation was

given. It was seen that the next generation of audio technology, such as immersive and object-based reproduction, offers a shift in experience towards immersive, interactive and personalised experiences. Before presenting methods for the evaluation of audio technology, considerations on the term "quality" in a general sense were given. It was seen that quality is the outcome of an individual's comparison and judgement process. As experienced quality is delimited in time, space and character and is unique to the user, relevant information about experienced quality can only be obtained on a descriptive level from the user. Probing this experienced quality is the aim of audio quality evaluation. Audio quality evaluation principles such as validity and assessor categorisation were then discussed and a range of audio quality evaluation techniques were presented. After a review of the state of the art of next generation audio technology evaluation, these audio quality evaluation techniques were discussed in light of the requirements for next generation audio assessment. Due to the potential of new experiences provided by next generation audio technology, and also the additional context and user factors that should be considered, the concept of quality of experience was identified as being highly relevant for the evaluation of next generation audio. It is the concept of QoE which is the topic of the following chapter.

3

QUALITY OF EXPERIENCE

3.1 INTRODUCTION

Whereas the previous chapter was concerned with audio technology and its evaluation, this chapter is predominantly concerned with quality of experience in general, and therefore quality evaluation in a more technology agnostic sense. Quality of experience is a measure that has grown and evolved over the last two decades. In the telecommunications field quality was traditionally related to quality of service (QoS), a technology-centric measure relating to service performance. QoE expands upon this by taking a more user-centric approach to measuring quality by taking factors into account such as expectations, perceptions and needs. QoE is not only relevant to the field of telecommunications. Multimedia services in general and other areas ranging from design to HCI are becoming increasingly interested in QoE. In some ways it could be seen that QoE is similar to User Experience (UX), and indeed, they do share some aspects. However, there are also key differences as discussed later in this chapter. In this section, the concept of QoE is introduced with discussions on definitions of QoE, Section 3.2, factors that influence QoE, Section 3.3, features of QoE, Section 3.4, and methods for the evaluation of QoE, Section 3.5. A discussion and summary can be found in sections 3.6 and 3.7 respectively.

3.2 DEFINITIONS OF QUALITY OF EXPERIENCE

In Section 2.3, definitions of the term "quality" and a description of the quality-formation process were given. The concept of quality of experience draws upon these ideas and can be defined as

> "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 fulfillment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the person's context, personality and current state"

(Möller and Raake, 2014, p. 19).[1] It should be re-emphasised that here we are talking about quality in a general, non-audio specific sense. The phrase "the person's evaluation of the fulfillment of his or her expectations and needs" can be seen as referring to the comparison and judgment segment of the quality-formation process (Figure 2.2), whereby the quality perception path is compared to the reference path. The phrase "in the light of the person's context, personality and current state" highlights the fact that QoE is dependent upon the context and the user and this is a distinguishing feature of QoE.

Whilst the above definition could be regarded as the most current, it is useful to consider previous definitions. A definition of QoE is given by ITU-T, (2008c, p. 2) as

> "the overall acceptability of an application or service, as perceived subjectively by the end user."

It is further noted that this includes the complete end-to-end system effects and it may be influenced by user expectations and content. According to Möller, (2010), a shortcoming of this definition is the inclusion of the term "acceptability". Acceptability is defined by Le Callet, Möller, and Perkis, (2013) as "the outcome of a decision which is partially based on the quality of experience". In comparison, the most current definition presented above refers to "delight" and "annoyance", thus suggesting a more hedonic view of QoE.

To help gain a greater understanding of the concept of QoE, it is beneficial to compare it with two closely related fields: quality of service and user experience.

3.2.1 *Quality of Experience Versus Quality of Service*

QoS is defined by ITU-T, (2008b, p. 3) as

> "the totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service."

When comparing this definition to that of QoE, some clear differences are seen. Firstly QoE has a wider scope; the above definition of QoS is clearly focussed on telecommunications services (although QoS is also applicable to fields such as computer networking) whereas QoE can be applied to a wider variety of fields. Secondly, contextual and user related factors are not adequately addressed by QoS. It could therefore

1 This definition is based on that given by Le Callet, Möller, and Perkis, (2013).

be seen that QoS takes a network-centric approach to quality whereas QoE takes a user-centric approach. Despite these differences, QoE is often highly dependent on QoS. The technical aspects of a systems performance can have a significant impact on certain dimensions of QoE.

3.2.2 *Quality of Experience Versus User Experience*

A comprehensive discussion on the similarities and differences between QoE and UX is given in (Möller and Raake, 2014, ch. 3), and it is this which we refer to here. Before comparing the two, it is first beneficial to present some definitions. According to the International Organization for Standardization (ISO, 2010), user experience can be defined as

> "a person's perceptions and responses that result from the use or anticipated use of a product, system or service."

To further clarify the attributes and characteristics of UX, Law et al., (2009) conducted a survey among researchers regarding their conceptions of UX. From this, UX is described as

> "dynamic, context-dependent and subjective, stemming from a broad range of potential benefits users may derive from a product"

Law et al., (2009, p. 722). Thus, UX is inherently subjective and individual (i.e. each experience is unique to the individual), as well as context dependent and dynamic. On the surface UX therefore sounds very similar to QoE, but there are differences.

The first difference between QoE and UX is their origins. QoE originates in the field of telecommunications and offers a shift in focus from the concept of QoS. UX originates in the field of HCI and offers a shift in focus from the concept of usability. Comparisons can be drawn between these paradigm shifts. Both QoS and usability are predominantly focussed on system and service performance related measures; QoS in terms of factors such as network performance, and usability in terms of factors such as users' efficiency and effectiveness in completing a certain task. Both QoE and UX expand on this performance-driven mindset by placing more emphasis on the human experience in general. The different origins of QoE and UX lead to a fundamental difference between the two. According to Roto et al., (2011), UX is not driven by technology but focuses on humans. QoE research on the other hand, is largely system and technology centred and is highly dependent on QoS.

This difference in theoretical basis leads to differences in measurement and evaluation techniques. The domain of UX has strong influences from domains such as psychology, sociology and ethnology, and as such has adopted a range of qualitative approaches from these disciplines. This in turn leads to human affects and emotions playing a prominent role in many UX studies. QoE measurements on the other hand are predominantly quantitative in nature. This difference in typical measurement techniques can be further explained by considering the philosophical grounding behind QoE and UX. Whereas QoE (to date) is heavily influenced by an empirical-positivist research paradigm, which results in experiments conducted in controlled environments in order to identify the impact of specific factors, UX is heavily influenced by an interpretive and constructivism-based research paradigm, which focuses on meaning and interpretation, and aims to gain a richer understanding of phenomena.

3.3 FACTORS INFLUENCING QUALITY OF EXPERIENCE

Quality of experience can be subject to a range of factors that influence the human experience. These "influence factors" (IFs) can be defined as

> "any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user"

(Le Callet, Möller, and Perkis, 2013, p. 11). As mentioned by Möller and Raake, (2014, p. 56), influence factors can therefore be considered as independent variables with the resulting QoE the dependent variable. Influence factors can be grouped into system, context and human IFs, although this distinction is not clear-cut. Due to the complex and interrelated nature of QoE IFs, these groups of influence factors often overlap and together have a mutual impact on QoE, as portrayed in Figure 3.1. In the following sections, these groups of influence factors are discussed in turn.

3.3.1 *System Influence Factors*

System influence factors (SIFs) can be defined as those that

> "refer to properties and characteristics that determine the technically produced quality of an application or service"

Figure 3.1: Factors influencing QoE can be grouped into human, system and context influence factors. As represented in the figure, these groups of influence factors often overlap and together have a mutual impact on QoE. Adapted from (Möller and Raake, 2014, p. 57).

(Le Callet, Möller, and Perkis, 2013, p. 11), based on (Jumisko-Pyykkö, 2011). These are related to aspects such as media capture, coding, transmission, storage, rendering, display and communication of information from content production to user. SIFs can further be divided into four subcategories: content-related, media-related, network-related and device-related.

Content-related SIFs refer to content type and content reliability. The content itself has a large impact on QoE and can include factors such as audio bandwidth, dynamic range, colour depth, texture, and spatial format (2D/3D). Media-related SIFs refer to media configuration factors such as encoding, resolution, sampling rate, frame rate and media synchronisation. Network-related SIFs refer to factors arising from data transmission over a network, such as bandwidth, delay and jitter. Finally, device-related SIFs refer to factors relating to end systems and devices, such as system and equipment specifications, device capabilities and provider specification capabilities. System IFs are well studied in the various branches of QoE research and many examples exist of studies investigating the IFs listed above.

System Influence Factors in Audio Evaluation

In the specific case of audio quality evaluation, studies on system IFs are also well represented. In the case of content-related SIFs, example studies include investigations on the effect of bandwidth limitation on audio quality (Zielinski, Rumsey, and Bech, 2003) and the studies discussed in Section 2.9 comparing spatial reproduction methods. Media-related SIFs are perhaps the most studied type of IF when it comes to audio evaluation. For instance, there are numerous studies on the effects of high resolution audio on perceived quality (Reiss, 2016), i.e. the effect of sampling frequency and bit-depth, as well as studies on the perceptual effects of codecs (coder-decoder), such as (Soulodre et al., 1998). Studies investigating network-related SIFs in the field of audio evaluation are most typically concerned with speech quality, for example see (Raake, 2006). Device-related SIFs are often interrelated with content-related SIFs in the field of audio evaluation as the type of content often depends on the playback device (e.g. with multichannel audio). Studies investigating factors of this type are also common, for example comparisons of loudspeakers (Toole, 1985) and spatial reproductions formats (Section 2.9).

3.3.2 *Context Influence Factors*

Context influence factors (CIFs) can be defined as

> "factors that embrace any situational property to describe the user's environment"

(Le Callet, Möller, and Perkis, 2013, p. 12), based on (Jumisko-Pyykkö, 2011). In a framework proposed by Jumisko-Pyykkö and Vainio, (2010) resulting from an extensive literature review of HCI studies, CIFs can be divided into physical, temporal, task, social, and technical and information factors. Economic factors are also included in (Le Callet, Möller, and Perkis, 2013). Furthermore, these factors can occur on different levels of magnitude (micro vs. macro), behaviour (static vs. dynamic) and patterns of occurrence (rhythmic vs. random).

The physical context describes the apparent features of a situation in which a given experience takes place. These include the spatial location (e.g. outdoor/indoor), functional space and place (e.g. zones in city areas), sensed environmental attributes (e.g. lighting, sound, haptics and weather), movements and mobility (e.g. sitting/standing), and artefacts (physical objects present).

The temporal context describes temporal aspects of a given experience. These include duration, time of day/week/year, actions related to time (e.g. hurrying/waiting), synchronism (i.e. synchronous actions such as talking on a phone versus asynchronous actions such as texting), frequency of use and properties in terms of before/-during/after an experience.

The task context describes the surrounding tasks in relation to a user's experience. These can include multitasking, interruptions and the type of task (e.g. work/entertainment, easy/difficult).

The social context is related to inter-personal relations existing during a given experience. These properties include persons present (e.g self/group, public/private, physical/virtual), interpersonal interactions (including all collaborative actions) and culture (including values, norms and attitudes of a certain culture).

The economic context includes factors such as costs, subscription type and brand of the application or system.

Lastly, the technical and information context describes relationships between the system of interest and other relevant services and systems. These include devices (e.g. interconnectivity of devices over Bluetooth), applications (e.g. availability of an application) and networks (e.g. availability of other networks). Also part of the technical context is interoperability between and across devices, applications and networks; informational artefacts, such as paper and pen; and mixed reality systems.

Context Influence Factors in Audio Evaluation

Compared to system influence factors, limited work has been done on the role of context influence factors in the field of audio evaluation. Beresford et al., (2006a)(2006b) compared audio quality ratings given in a listening room to those given in an automotive setting in a laboratory. However, no conclusions on the role of context could be made. Fiebig, (2015) presents a discussion on the influence of context effects on sound quality assessments with results from two case studies. Vacuum cleaner sounds were evaluated in both a home-environment and in a laboratory situation. For the home-environment the actual products were used and the subjects were requested to simultaneously watch television, whereas in the laboratory situation binaural recordings of the products were used. Users were less critical of the product sounds in the home-environment and this was linked to the differing levels of attention given in the two situations. Additionally, kettle noises were assessed with and without visual information, but no effect of visual input on the sound assessment was found. With regards to

the influence of contextual factors on the perception of spatial audio, previous work has focussed on the plausibility of binaural recordings with respect to listening context. Werner and Klein, (2014) investigated how context influences the quality parameters "externalisation" and "direction of auditory event" for binaural reproduction. When participants could see the listening room they were in, including visual clues of dummy speaker locations, a greater sense of externalisation was reported. Also, when there was a divergence between the listening room and synthesised (audible) room, a lesser sense of externalisation was reported.

Several approaches have been taken to adapt audio content to the context, specifically environmental noise, to improve the listening experience. For example, Reis, Carriço, and Duarte, (2009) developed and evaluated a prototype system that utilises inbuilt microphones on regular mobile devices to adjust the volume of communication and media applications according to different aspects of context, whilst considering user preferences. As well as using environmental noise as a contextual variable, they also considered users' hearing capabilities. A slightly different approach was taken by Mason et al., (2015). They developed a system known as "personalised compression" that adapts the dynamic range of the audio being played according to the environmental noise in the listening situation. To investigate the relation between environmental noise and preferred loudness range, Kean, Johnson, and Sheffield, (2015) conducted an experiment in which participants were asked to adjust the loudness of audio content whilst listening in the presence of reproduced environmental noise, for both headphone and loudspeaker listening. By comparing loudness changes of the content to gain changes made by the listeners, they produced compensation slopes, which describe the preferred level of loudness compensation for different conditions.

3.3.3 *Human Influence Factors*

Human influence factors (HIFs) can be defined as

> "any variant or invariant property or characteristic of a human user. The characteristic can describe the demographic and socio-economic background, the physical and mental constitution, or the user's emotional state"

(Le Callet, Möller, and Perkis, 2013, p. 11). The subjectivity of HIFs makes them complex to study, as well as the fact that they can be strongly interrelated with other types of IFs.

HIFs can influence the perceptual process at two levels (Jumisko-Pyykkö, Häkkinen, and Nyman, 2007). At the early sensory or low-level processing level, properties related to the physical, emotional and mental constitution of the user are found. These properties may be dispositional such as auditory acuity, gender and age (Strohmeier, Jumisko-Pyykkö, and Reiter, 2010), or dynamic such as emotions, mood, personality, motivation and attention (Reiter and De Moor, 2012). Likewise, at the level of higher-level cognitive processing where interpretation and judgement is important, both invariant and dynamic properties can be found. Invariant properties could include socio-economic situation, education background and values, whereas dynamic properties could include expectations, needs, knowledge, previous experiences and emotions (Geerts et al., 2010; Wechsung et al., 2011).

Previous studies in the field of QoE have investigated a range of these HIFs. For example, Jumisko-Pyykkö and Häkkinen, (2008) investigated the impact of psychographic variables on the consumer-oriented quality assessment of mobile television. The studied variables were age, gender, education, professionalism, television consumption, experiences of different digital video qualities, and attitude towards technology. The results showed that quality evaluations were affected by almost all background factors. In a study by Wechsung et al., (2011), it was shown that attitudes and mood are related to quality perceptions, yet no link was found between personality traits and perceived quality. Other studies include those looking at the influence of mood and emotions (Arndt et al., 2012; Rainer et al., 2012; Reiter and De Moor, 2012), motivation (Ryan and Deci, 2000) and expectations (Sackl, Schatz, and Raake, 2017; Sackl et al., 2012; Staelens et al., 2012) on QoE.

Human Influence Factors in Audio Evaluation

With the exception of previous experiences and prior knowledge, the study of human influence factors in the field of audio evaluation is much more limited. Previous experiences and prior knowledge are typically used to distinguish between expert and naïve listeners, as discussed in Section 2.4.3. Quintero and Raake, (2012) however, investigated how factors beyond the level of prior knowledge of users affects perception of quality in the context of speech quality evaluation. Users were classified into six groups according to their demographic characteristics, their attitude towards adopting new technologies and socio-economic information. Significantly different quality ratings between these groups were found. Other studies include those looking at the influence of cultural backgrounds on timbre preferences (Kim, Bakker, and Ikeda,

Table 3.1: Overview and examples of the various forms of influence factors. Modified from (Möller and Raake, 2014, p. 68).

IF	Type	Examples
System	Content-related	Audio bandwidth, dynamic range, colour depth, texture, spatial format (2D/3D)
	Media-related	Encoding, resolution, sampling rate, frame rate, synchronisation
	Network-related	Bandwidth, delay, jitter, error-rate
	Device-related	Display resolution, colours, brightness, audio channel count, loudspeaker properties
Context	Physical	Location and space, environmental attributes, motion
	Temporal	Time, duration, frequency of use
	Social	Persons present, interpersonal actions, culture
	Economic	Costs, subscription type, brand
	Task	Multitasking, interruptions, task type
	Techincal	Compatibility, interoperability, additional informational artefacts
Human	Low-level	Sensorial acuity, gender, age, lower-order emotions, mood, personality, motivation, attention
	High-level	Socio-economic situation, education, values, expectations, needs, knowledge, previous experiences, emotions

2016), the influence of listeners' experience, age, and culture on headphone sound quality preferences (Olive, Welti, and McMullin, 2014) and various studies looking at the impact of language on quality perception of speech (Ebem et al., 2011; Schinkel-Bielefeld et al., 2017).

3.3.4 *Summary*

As highlighted above, QoE (both in a general and audio specific sense) can be subject to a wide range of influence factors that can often be complex and strongly interrelated. A summary of these is given in Table 3.1. In terms of audio evaluation, SIFs are by far the most widely studied, whether it be investigations on codecs, spatial reproduction methods or loudspeaker properties. Studies on CIFs and HIFs, however, are much rarer. Table 3.1 illustrates the breadth of factors that CIFs and HIFs incorporate

and many of these are yet to be studied in the context of audio evaluation. There is therefore great potential for improving the quality of experience of audio systems and services by taking a QoE mindset for its evaluation, which considers system, context and human influence factors.

3.4 FEATURES OF QUALITY OF EXPERIENCE

Features of quality of experience relate to how the influence factors described in the previous section are perceived by the user. The characteristics of an individual's experience can be analysed by decomposing the experience into such "quality features", a process which is necessary in understanding why a given experience is experienced in the way it is. A QoE feature can be defined as

> "a perceivable, recognized and nameable characteristic of the individual's experience of a service which contributes to its quality"

(Le Callet, Möller, and Perkis, 2013, p. 13), following definitions by Jekosch, (2005). Again, here quality is referred to in a general, non-audio specific sense.

Referring back to the quality-formation process presented in Section 2.3.2, perceived quality features result from the sensing of a physical source and the subsequent reflection on this perceived sensation. This process is situation and context dependent and the relationship between physical "quality elements" and perceived "quality features" is multivariate. As such, quality features can be represented in a multidimensional perceptual space and therefore analysed using multidimensional analysis. Furthermore, the context dependent nature of quality features means that empirical analysis of quality features often only reveal features that are perceivable in the respective context (Möller and Raake, 2014, p. 74).

Quality features can be categorised into five levels, (Möller and Raake, 2014, pp. 78 - 80) based on (Le Callet, Möller, and Perkis, 2013):

- *Level of direct perception* - these are QoE features that relate to the perceptual information created immediately and spontaneously during the media consumption. Examples related to audio could include localisation, timbre and envelopment.
- *Level of action* - this level relates to the human perception of ones own actions. Examples in audio include the perception of ones own voice in speech services (e.g. echo) and in video services include involvement and immersion, the perception of space and the perception of ones own motions.

- *Level of interaction* - this level deals with human-to-human and human-to-machine interaction and includes features such as responsiveness, naturalness of interaction, communication efficiency and conversation effectiveness.
- *Level of the usage situation of the service* - this level relates to the physical and social situation. Examples include accessibility and stability during usage.
- *Level of service* - this relates to usage of service beyond a particular instance. Examples include aesthetic feeling, usability, usefulness, joy and ease of use.

From this classification, it is apparent that the majority of quality features extracted for the purpose of audio evaluation (excluding speech services) are at the level of direct perception.

The extraction of quality features is possible by a range of methods, the aim being to identify and quantify the features relevant to a given experience. Section 2.6 presented various attribute-based methods for audio evaluation and all of these could be regarded as feature extraction methods. Outside of audio evaluation, other methods such as interview-based techniques (Jumisko-Pyykkö, 2011) and Open Profiling of Quality (Strohmeier, Jumisko-Pyykkö, and Kunze, 2010) (discussed further in Section 3.5) have been used to create general sets of QoE features. Regression techniques can then be used to determine the relevance of the identified QoE features for quality preferences (such as external preference mapping). Moreover, relationships between perceptual features and quality metrics are often sought, as discussed in the context of sound quality models in Section 2.8.

3.5 METHODS FOR MULTIMEDIA QUALITY OF EXPERIENCE EVALUATION

Many of the concepts presented on the topic of audio quality evaluation in Section 2.4, such as validity and assessor categorisation, are also relevant for the evaluation of multimedia QoE in general. As many of the methods are also similar, only a brief overview of the types of methods employed for the evaluation of multimedia QoE are presented in this section. For the evaluation of audio quality, we saw that methods can be grouped into those that use global judgments, those that are attribute-based, and those that combine both global and attribute judgments. For the case of multimedia QoE evaluation, we shall group methods into the related but not synonymous groups of quantitative methods, qualitative methods and mixed methods.

3.5.1 *Quantitative Methods*

In a review of mixed methods research, Johnson and Onwuegbuzie, (2004, p. 18) describe quantitative research in the following manner:

> "the major characteristics of traditional *quantitative* research are a focus on deduction, confirmation, theory/hypothesis testing, explanation, prediction, standardized data collection, and statistical analysis."

In terms of the audio evaluation techniques previously presented, all global judgment methods and those attribute-based methods that do not include some form of elicitation could be considered as quantitative. Similar techniques to these are also found for multimedia evaluation, e.g. (ITU-R, 2012b) and (ITU-T, 2008a), in which quality is assessed globally or with a pre-determined set of attributes. Such methods are commonly used in QoE evaluation due to its QoS origins. This illustrates that whilst a paradigm shift occurred from QoS to QoE, the dominant evaluation scale (MOS) remained the same (Möller and Raake, 2014, p. 47). As QoE is multidimensional and is dependent upon the context and the user, it is unlikely that such methods by themselves are sufficient for in-depth QoE evaluation.

Physiological Measures

As well as measuring QoE directly (i.e. asking for a response), QoE can be measured indirectly. Physiological measures are one tool to measure QoE indirectly and are used with the aim of objectively measuring emotional responses. Examples of physiological measures include heart rate, galvanic skin response, eye tracking, electromyograms (EMG) for detecting muscle (e.g. facial) activity and electroencephalograms (EEG) for measuring brain activity. For a comprehensive review of studies investigating physiological measures and emotion, see Kreibig, (2010). Whilst physiological measures have been proven to be valuable in the assessment of QoE, they do have limitations, including inherent physiological differences between humans resulting in noisy and possibly erroneous data, often expensive and complex experimental setups, and intrusiveness of measurement techniques (e.g. attaching sensors to participants) causing changes in natural behaviour and a reduced external validity (Engelke et al., 2017).

3.5.2 *Qualitative Methods*

Qualitative research, on the other hand, has the following characteristics:

> "the major characteristics of traditional *qualitative* research are induction, discovery, exploration, theory/hypothesis generation, the researcher as the primary 'instrument' of data collection, and qualitative analysis"

(Johnson and Onwuegbuzie, 2004, p. 18). In terms of the audio evaluation techniques previously presented, the methods with a vocabulary-based elicitation stage could be considered to contain qualitative aspects. For multimedia QoE evaluation, qualitative methods generally take one of two approaches: interviews and sensory evaluation.

With interview-based methods, participants explicitly describe the characteristics of the stimuli by means of free-description or stimuli-assisted tasks. Semi-structured interviews, i.e. interviews containing a pre-determined set of open questions, are a commonly used qualitative method as they can provide detail, depth, and the perspective of the participant while at the same time allowing hypothesis testing and the quantitative analysis of interview responses (Leech, 2002). An example of interview-based methods being used for multimedia QoE assessment is presented by Jumisko-Pyykkö, (2011), in which interview-based techniques are used for the assessment of mobile television. It should be noted however, that in this example the interviews are used in conjunction with quantitative methods.

Sensory evaluation methods that contain attribute elicitation, such as those presented in Section 2.6, can also be used for multimedia evaluation. As these have previously been discussed, they shall not be listed again here. One example of a sensory evaluation method for the assessment of multimedia is the rapid perceptual image description (RaPID) method (Bech et al., 1996), which is a descriptive analysis method for assessing the image quality of televisions.

Another group of methods that should be mentioned are those related to observation of behaviour and interaction. Such methods are commonly found in UX and HCI research, but are less well represented in the context of QoE evaluation. Methods of this type range from purely qualitative, for instance ethnography, to those with quantitative aspects, such as those that quantify interactions with interfaces. As an example, in the context of mobile video streaming, Huang, Zhou, and Du, (2014) analysed user behaviour including pausing, fast-forwarding, switching video resolution and quitting in order to evaluate QoE. Due to the nature of next generation audio, methods such as these may prove an interesting alternative for its evaluation.

3.5.3 *Mixed Methods*

According to Johnson and Onwuegbuzie, (2004, p. 17), mixed methods research is the "third-wave" or third research movement and makes use of the pragmatic method and system of philosophy. They define mixed methods research as

> "the class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study".

Combing quantitative and qualitative research methods is said to be able to draw from the strengths of each whilst compensating for any weakness, thus increasing insight, understanding and generalisability of results. In terms of the audio evaluation techniques previously presented, mixed methods studies include those that combine quantitative rating tasks with qualitative tasks, such as attribute elicitation, in the same study.

For multimedia QoE evaluation, Jumisko-Pyykkö, (2011) combined interview-based techniques with quantitative methods, as mentioned above. More specifically, quantitative quality evaluation tasks were followed by post-task interviews in order to explore experienced quality factors for audiovisual quality of mobile television. The advantages of such a method are that descriptive interviews can be conducted quickly and can be adapted for various contexts (e.g. in (Jumisko-Pyykkö and Utriainen, 2010)). However, the qualitative data are based on all stimuli, meaning that analysis of single stimuli is limited. Another mixed methods approach to multimedia evaluation is Open Profiling of Quality (OPQ) (Strohmeier, Jumisko-Pyykkö, and Kunze, 2010). OPQ consists of three primary sections; a psychoperceptual evaluation stage aims to evaluate the degree of overall quality, a sensory profiling stage aims to explore the profiles of the overall quality by means of individual vocabulary elicitation and attribute rating, and finally an external preference mapping stage aims to study the relationship between the overall quality and the quality profiles. The method, which is intended for naïve assessors, was originally developed for the quality evaluation of visual and audiovisual systems, specifically mobile 3D television and video, and has been used in both laboratory and field situations (Strohmeier, 2012; Strohmeier, Jumisko-Pyykkö, and Eulenberg, 2011; Strohmeier et al., 2011).

3.6 DISCUSSION

The above sections have been predominantly concerned with concepts related to QoE in general. The topics presented are however highly relevant for the specific case of next generation audio quality assessment. In order to improve the experienced quality of next generation audio, a QoE mindset should be taken that considers system, context and human factors. As one aspect of next generation audio is personalisation and adaptation, context and human factors are central to evaluating and optimising the experience. System influence factors are widely studied in the field of audio quality evaluation, but by studying these alone important information about how situational and user specific factors influence the listening experience may be overlooked. It is therefore beneficial for the audio community to consider the theory of QoE when evaluating next generation audio. However, it should be kept in mind that the theory of QoE is not necessarily reflected in evaluation methods typically used for QoE evaluation. Both the QoE and audio communities could benefit from utilising qualitative aspects in order to find the relevant quality features that influence preference in the given context and for the given user. Mixed methods such as Open Profiling of Quality and other methods that combine preference ratings with sensory profiling therefore show potential for QoE evaluation. Indeed, such methods are further explored in the following chapters of this thesis.

It is with the above discussion in mind, as well as the literature presented in the previous chapter, that the research questions outlined in Section 1.2 were formed. To restate these, the general aims of this research are i) to explore the role of system, context and human influence factors on the QoE of next generation audio, and ii) to investigate suitable methods and approaches for the evaluation of the QoE of next generation audio, with respect to its various influence factors.

3.7 SUMMARY

In this chapter an overview of the concept of quality of experience has been presented. It was seen that QoE is a measure that has grown and evolved over the last two decades and represents a divergence from the more typical signal-based measures of quality, known as quality of service. QoE expands upon signal-based measures of quality by taking factors into account such as the user's context, personality and current state. QoE can be subject to a range of factors that influence the human experience

and these influence factors were discussed. It was seen that IFs can be categorised as system, context and human IFs, although this distinction is not clear cut; these groups of influence factors often overlap and together have a mutual impact on QoE. With regards to audio quality evaluation, system IFs are extensively studied, yet studies related to context and human IFs are much rarer. Following the discussion on influence factors, quality features were discussed. Such features refer to the characteristics of an individual's experience and can be represented in a multidimensional perceptual space. To identify and quantify the features relevant to a given experience, a range of methods can be used. These include attribute-based methods, such as those used for audio evaluation, interview-based techniques and also methods such as Open Profiling of Quality. Methods for multimedia quality of experience evaluation was the final topic of this chapter, in which quantitative, qualitative and mixed methods were considered. Quantitative methods that use global judgments, such as those that rate mean opinion score, are frequently used in QoE assessments due to its QoS origins. It is unlikely that such methods alone are sufficient for QoE evaluation as QoE is multidimensional and is dependent upon the context and the user. Mixed methods offer the potential to relate overall quality or preference with the relevant quality features for the context in question, and are therefore potentially more useful for the evaluation of QoE.

Part I

SYSTEM

4

A SUBJECTIVE COMPARISON OF DISCRETE SURROUND SOUND AND SOUNDBAR TECHNOLOGY

4.1 INTRODUCTION

In the first of the studies presented in this thesis, the role of system influence factors on the quality of experience of next generation audio are explored. As previously discussed, system IFs relate to aspects such as media capture, coding, transmission, storage, rendering, display and communication of information from content production to user. In the case of next generation audio technology, reproduction methods are one system IF that can potentially have a large influence on quality of experience. In the study presented here, the specific case of soundbar reproduction is used to explore the impact of system IFs on the quality of experience of next generation audio, as well as to explore suitable methods by which to do so.

Soundbars, as was seen in Section 2.2.4, offer a more convenient approach to deliver a surround audio experience than traditional discrete speaker setups. Such technology is generally advertised as being able to deliver a good spatial impression from a single enclosure of loudspeakers and this therefore makes soundbars relatively simple to setup and more practical for domestic environments. Soundbar technology has the potential to progress from providing surround experiences to providing full immersive experiences and the technology could therefore enable immersive audio reproduction in the home for a large proportion of users.

The potential for soundbars to deliver immersive experiences to the user, which is a characteristic of next generation audio, is just one reason why evaluating soundbar technology is a beneficial case study to gain insight into the role of system influence factors on the QoE of next generation audio in general. Due to their design, namely their slim profile and limited separation of drivers, the experience provided by soundbars is likely to be perceptually different than that provided by other reproduction technologies (such as discrete systems) due to a broad range of quality features. It is expected that both timbral and spatial quality features will be pertinent to the QoE

provided by soundbars and therefore the method used to evaluate such technology should reflect the multidimensionality of perceived quality. Moreover, it could be argued that there is no high quality reference to compare soundbars against as they could provide perceptually different listening experiences than other immersive reproduction technologies, but be of equal "quality". Both of these factors mean that any method designed to evaluate soundbar listening will likely be applicable to a range of other next generation audio technologies and system IFs.

Furthermore, whereas the quality of other next generation audio reproduction systems has been the focus of previous studies (as described in Section 2.9.1), prior to this research there had only been limited published work on the perceptual evaluation of soundbar technology. Moulin, Nicol, and Gros, (2012) compared a discrete surround system, headphones and a soundbar in the context of an audiovisual scenario. The soundbar was rated as having a lesser degree of sound spatialisation than the two other systems although the test took place in an acoustically treated room - possibly influencing the performance of the soundbar, as discussed further on in this chapter. Additionally, the term "degree of sound spatialisation" was used which could be seen to encompass various aspects of the sound, limiting the scope of the results. It is therefore necessary to provide a more thorough perceptual evaluation of soundbar technology in order to assess the experience currently provided by this technology.

On the other hand, using soundbar listening as a case study for investigating system IFs in the context of next generation audio does also have some limitations. Compared to other next generation reproduction systems such as discrete setups, soundbars potentially have a greater variance in provided experience between the different products on the market, due to the range of product quality and technology used. This, coupled with the fact that there is generally a lack of specific technical information about soundbars from manufacturers, means that quality evaluations will be less generalisable than the case for other reproduction methods. On top of this, context effects will likely influence the results of soundbar quality evaluations more than for other reproduction methods, again decreasing the generalisability of results. Despite these limitations, an evaluation of the QoE of soundbar listening will still give valuable insights into the current state of soundbar technology.

Considering the above points, it is deemed a useful contribution to consider soundbar listening as a case study to explore system IFs with regards to next generation audio, both on a specific level for evaluating the QoE of soundbars, as well as on a

more general level of investigating suitable methods that could be applicable to the evaluation of other next generation audio system IFs.

The specific aim of the study presented here is to assess the effectiveness of soundbar technology by subjectively comparing a discrete surround system, a discrete stereo system and two soundbars for a range of content material. Furthermore, two groups of participants grouped according to listening experience are compared as it has previously been shown that experienced and naïve assessors base their preferences of multichannel audio on different aspects of quality, see Section 2.4.3.

To evaluate the experience provided by next generation audio technology, it was argued in Section 2.10 that methods that combine preference ratings with sensory profiling are suitable as such methods offer the possibility of determining the relationships between listener preference and the attributes that play a role in the formation of preference. In Section 3.5.3 a mixed methods approach called Open Profiling of Quality was introduced that combines overall quality ratings with sensory profiling in the same method. Although this method was originally intended for the evaluation of visual and audiovisual systems, it is also suitable for the evaluation of audio technology for the following reasons. First and foremost, both visual technology and audio technology reproduce stimuli that are typically heterogeneous and multidimensional in character. Furthermore, both forms of technology have the potential to deliver novel experiences to the user, whether it be through 3D video or binaural audio, and both are used in a wide range of contexts. Compared to the other methods that combine global judgments with attribute ratings presented in Section 2.7, OPQ has the following features.

i) It is an individual vocabulary technique, meaning that each participant develops and employs their own attribute list for further rating. An advantage of such methods is that, compared to non-individual methods, participants may be able to better relate to the attributes being used (Berg, 2006).

ii) It is suitable for naïve listeners, as well as experienced listeners.

iii) It is relatively time-efficient compared to other similar methods, as the adaption described below only requires two sessions and does not include a panel discussion session.

For these reasons, it was decided that OPQ would be an ideal method for the purposes of this experiment. In addition to evaluating the experience provided by soundbar

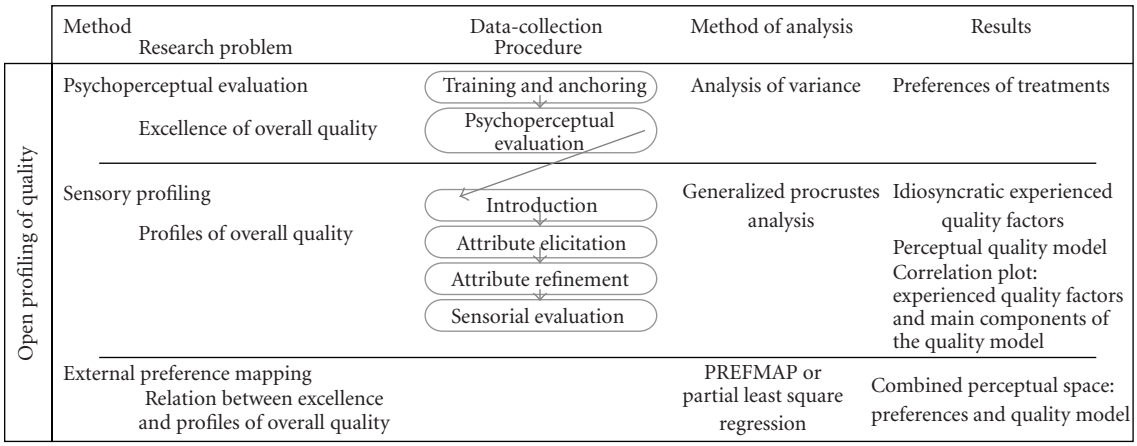


Figure 4.1: Overview of the original OPQ structure (Strohmeier, Jumisko-Pyykkö, and Kunze, 2010).

technology, this experiment therefore also had the aim of assessing and adapting the OPQ method for the application of comparing audio reproduction systems.

4.2 METHODOLOGY

In this section, a description of the OPQ method is given. Differences between the original implementation (Strohmeier, Jumisko-Pyykkö, and Kunze, 2010) and the adapted implementation for the comparison of audio reproduction systems presented here are highlighted, although these differences are discussed more thoroughly in Section 4.5.

4.2.1 *Structure*

As previously mentioned in Section 3.5.3, OPQ consists of three primary sections: a psychoperceptual evaluation stage, a sensory profiling stage and an external preference mapping stage, see Figure 4.1. In the original implementation of the method, the psychoperceptual evaluation and sensory profiling were conducted in different sessions. This was modified in this study, as shown in Figure 4.2. The purpose of this restructuring was to reduce the duration of the experiment and to aid in the elicitation of attributes that led to listener preference, as discussed in the following sections. Session 1 had a total duration of 90-120 minutes and session 2 had a total duration of 60-90 minutes. These were completed on separate days and within five days.

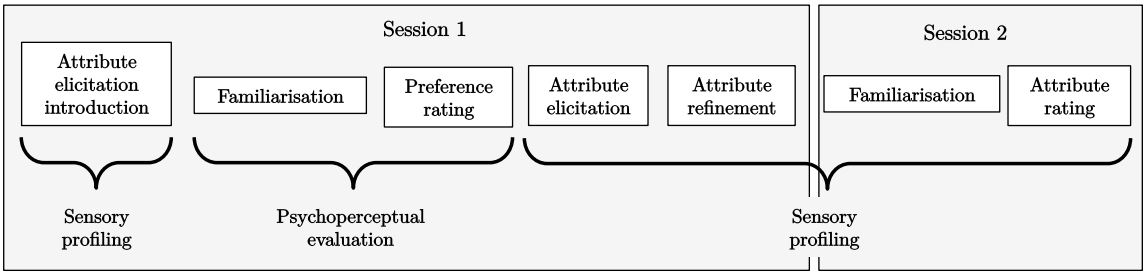


Figure 4.2: Overview of the adapted OPQ structure employed for this study.

4.2.2 *Attribute Elicitation Introduction*

After reading an information sheet and filling in a short demographic survey, the first task presented to the participants was an introductory verbal exercise on attribute elicitation. The purpose of this task was to familiarise participants with the attribute elicitation process and the kind of words that may be used to describe sensory differences and similarities between two objects. The question posed to the participants was "Imagine a basket full of apples. What kind of attributes, properties or factors can you use to describe similarities and differences of two randomly picked?". The researcher could help the participants find attributes but never suggested specific examples. It was decided not to use an introduction exercise related to audio so as not to bias the future auditory elicitation task.

4.2.3 *Familiarisation*

Before beginning the rating stages it is important to familiarise participants with both the stimuli that will appear in the experiment and the user interface. By familiarising participants with the scope and range of stimuli that will be used, participants will be able to use the rating scales more effectively which will in turn help reduce scale related bias (Bech and Zacharov, 2006). A simple method of achieving this is to allow participants to select and play a number of samples that span the range of qualities to be evaluated. In this study, the different quality levels relate to different reproduction systems, although in other studies these quality levels could relate to a range of other system IFs.

A familiarisation page using the listening experiment software (see Section 4.3.7) was presented to the participants after the attribute elicitation introduction. On this page were two rows of four excerpts - "A", "B", "C" and "D", corresponding to the

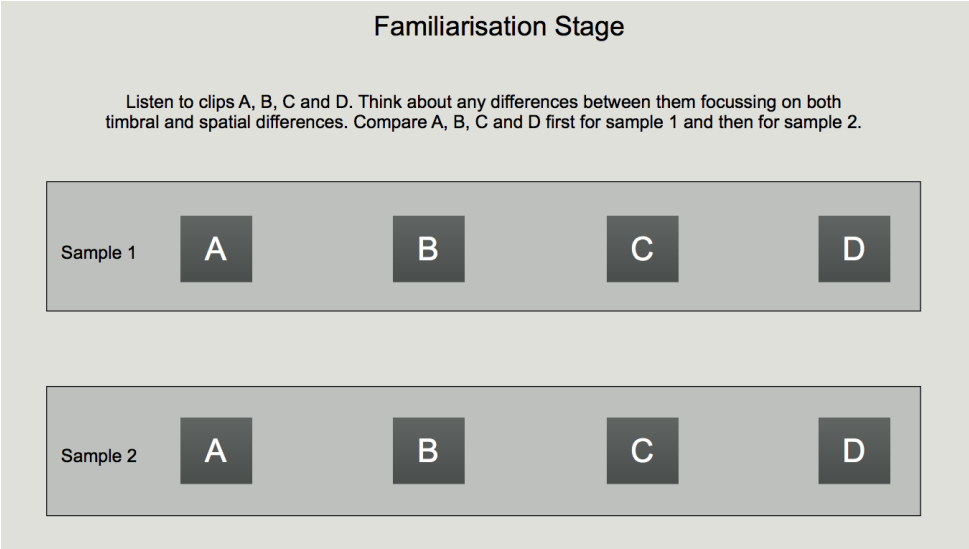


Figure 4.3: Familiarisation interface.

four reproduction methods for two different familiarisation stimuli, Figure 4.3. The participants were instructed to listen through the two series of four excerpts thinking about the differences between them, both in terms of timbral and spatial aspects. For naïve listeners the distinction between timbral and spatial aspects was explained and discussed. The simple explanation given was that spatial aspects refer to the perceived position of the sounds heard, whereas timbral aspects encompass everything else. This explanation could be seen as a mild form of training, resulting in participants paying more attention to spatial aspects than they otherwise would initially. On the one hand this could result in more consistent judgments as participants are likely to pay attention to both aspects of quality from the outset. However, it could also be seen as a negative priming effect if the response desired is that from a totally untrained listener. Participants were also encouraged to think about which clip they preferred and why. It was stated to the participants that the degree of difference between the audio clips presented here was representative of the rest of the experiment.

4.2.4 *Preference Rating and Attribute Elicitation*

In the original implementation of OPQ, Absolute Category Rating (ACR) with single stimuli was employed for the psychoperceptual evaluation stage. In this study, this was adapted to a paired comparison, preference rating method. It is important to note that ACR is used for the rating of overall quality, as previously discussed in

Section 2.5 (Figure 2.7a), which is different to "preference". When comparing spatial sound systems without introduced degradations, it could be the case that quality is perceived as equally high for all systems, even though participants may have certain preferences. For this reason, preference ratings were deemed to be more suitable than overall quality ratings for the case of comparing reproduction systems. Advantages of paired comparisons is that they are powerful when comparing systems with small differences whilst still being simple for untrained participants, although this is at a cost of an increased experiment duration. This increased power could be particularly important for studies investigating other system IFs, where the differences between quality levels could be more subtle.

A full paired comparison method was employed which resulted in 36 pairs to be rated (six pairs of systems for six content items). Each pair was the same content item played over two different systems. For the preference rating, the question posed to the participants was "compare clips 'A' and 'B' in terms of which you would prefer to listen to in your home. Listen for both timbral and spatial differences between the clips". The participants could then make a rating on a scale indicating preference to either "A" or "B" (see Section 4.3.7 for more details on the interface). Below the rating scale was an empty text box with the instructions "list any differences between A and B that led to this decision. Include both timbral and spatial differences". Here, the participants listed adjectives or phrases describing the differences that they heard. Participants took a 10-15 minute break half way through this section so as to limit listener fatigue.

Whereas attribute elicitation was conducted in a separate session to preference rating in the original implementation of the method, here attribute elicitation was carried out simultaneously with preference rating. The aim of this was to encourage participants to list attributes that were directly related to their given preference ratings. Moreover, with this format there were two sections that involved listening instead of three with the original format, possibly reducing listener fatigue. In the original implementation Strohmeier, (2011) cites Faye et al., (2006) when discussing the order of the psychoperceptual evaluation and sensory profiling tasks. Faye et al., (2006) advise that hedonic tasks should be completed before sensory profiling tasks as this results in the preference ratings being completely "clear of influence". However, in their study in which two groups of participants were used to explore the role of task order on results, no influence was found. This suggests that combining the preference rating

and attribute elicitation stages, as presented here, should have a minimal negative influence on the results.

4.2.5 *Attribute Refinement*

For accurate profiling it is necessary to refine the list of attributes that participants develop. After the preference rating session, participants were presented with an on-screen text file containing all of the descriptions that were written into the text box during the preference rating and elicitation process. In order to refine this list, participants were first asked to pick out unique attributes and to write these into a new text file. This included grouping opposite terms together so that either the positive or negative remained. To ensure that the remaining attributes were unique, i.e. did not cover the same aspect of quality, participants were asked to explain the difference between attributes that sounded similar. Secondly, if there were more than eight attributes at this stage, participants were asked to choose the eight attributes that were the most influential on their preference ratings. By limiting the number of attributes to eight, the attribute rating session was kept to an acceptable duration. Some participants at this stage modified the attributes to better explain their intended meanings. It should be noted that the researcher helped with this process although never suggested specific words. To conclude the attribute refinement process, participants were asked to describe the final list of attributes to the researcher including clarifying what more or less of each attribute meant.

4.2.6 *Attribute Rating*

As this stage took place on a different day, the familiarisation stage was repeated prior to the attribute rating. Participants were handed a list of their developed attributes and were asked to think about the differences between the familiarisation clips with respect to the listed attributes.

The aim of the attribute rating stage is to quantify the strength of the developed attributes for each stimuli. As with the preference ratings, a full paired comparison method was used to achieve this. Each participant used an individualised interface with rating scales corresponding to the developed attributes from the previous stage. They were instructed to rate which clip, "A" or "B", had more of the listed attributes

on a rating scale ranging from "A has much more" to "B has much more". Again, 36 comparisons were made with a 10-15 minute break half way through.

The paired comparison method used is in contrast to a single stimulus method described in the original implementation. Rating all attributes in the same trial for each paired comparison has the advantage of reducing the test duration compared to rating attributes in succession, although it should be noted that one disadvantage could be inter-attribute correlations, i.e. a general preference for a stimulus might show as increased attribute ratings in favour of that stimulus.

4.3 EXPERIMENTAL DESIGN AND SETUP

4.3.1 *Reproduction Systems*

A variety of technology exists when it comes to audio reproduction via soundbars; the exact details of which are often not publicised by manufacturers or discussed in academic literature. However, it is possible to make several distinctions between the various technologies available. The first is a distinction between stereo soundbars, which down-mix additional channels to stereo, and soundbars which can process surround channels individually. It is the latter of these that are of interest in this study as they can be considered as next generation audio devices with the potential to deliver immersive experiences. Although devices of this type on the market can, at the time of this study, only process horizontal-only surround content, in the future such devices are likely to be able to process with-height, immersive content.

Another distinction is made between the methods that soundbars use to reproduce the spatial characteristics of audio content. It appears that one group of soundbars focuses on beamforming technology, by which sound is reflected off the walls of the room in order to replicate discrete speaker locations, as discussed by Hooley, (2006). Such soundbars can be identified by their large number of drivers, typically many more than the number of channels being reproduced. Another group does not use beamforming but rather focuses on psychoacoustic filtering to create sound zones around the listener. These soundbars have fewer drivers, typically around the same number as the number of channels being reproduced. It is with this in mind that two soundbars were chosen for this study. Soundbar 1 (s1), a Focal Dimension, predominantly uses filtering to achieve a surround sound experience (Focal, 2014). Soundbar 2 (s2) on the other hand, a Yamaha YSP-4300 "digital sound projector", predominantly

uses beamforming (Yamaha, 2012). To assess the effectiveness of soundbar technology, these two soundbars were compared to a discrete surround system and also a discrete stereo system.

As mentioned in the introduction of this chapter, comparing such different reproduction systems will likely elicit a broad range of attributes that are relevant for the evaluation of next generation audio reproduction systems. As well as using fundamentally different techniques to spatialise the content, there is also a large difference in the size of the drivers and cabinets between the different systems. As a result, both spatial and timbral differences should be apparent between the systems.

4.3.2 *Stimuli*

Six audio excerpts in the 5.0-channel format were used for the main sections of this experiment with an additional two being used for a familiarisation stage. These spanned a range of genres (ambient sound, pop music, classical music, radio drama, documentary and film) in order to cover the range of material likely to be experienced by users of the systems under study. The excerpts chosen contained both foreground-foreground (F-F) characteristics and foreground-background (F-B) characteristics (Zielinski, Rumsey, and Bech, 2002) so that the excerpts had varying degrees of spatial information. A basic measure of how much surround information each stimuli contained was quantified with two ratios: a root mean square (RMS) frontal-surround ratio ($R_{FS,rms}$) and a peak frontal-surround ratio ($R_{FS,peak}$). The RMS frontal-surround ratio was calculated by dividing the sum of RMS values for the L, R and C channels with the sum of RMS values for the LS and RS channels and converting to dB, i.e.

$$R_{FS,rms} = 10 \log_{10} \left(\frac{L_{rms} + R_{rms} + C_{rms}}{LS_{rms} + RS_{rms}} \right). \tag{1}$$

The peak frontal-surround ratio was calculated in the same way but using peak values. The higher the ratio, the less surround information a sample has. Details of the stimuli are presented in Table 4.1.

The excerpts ranged in duration from 10 to 15 seconds and included an added 1.5 seconds of silence at the beginning of each clip. This silence ensured that the reproduction systems were not identifiable by their fade-in characteristics once they started to receive a signal. For each of the stimuli, stereo down-mixes were created in accordance with ITU-R BS.775 (ITU-R, 2012a) for playback over the discrete stereo

Table 4.1: Overview of content items used including spatial categorisation and frontal-surround ratios. Starred items were used in the familiarisation stage only.

Item Genre	Title	Categorisation	R_{FS} (dB) rms	R_{FS} (dB) peak	Description
Documentary	Africa	F-B	8.2	8.8	BBC nature documentary. Music and effects in the stereo channels, dialogue and effects in the centre channel, effects and percussion in the surround channels.
Ambient sound	Applause from Last Night of the Proms	F-F	0.0	0.0	Crowd from BBC live classical music recording. Applause in all channels.
Pop music	Lady Gaga - Pokerface	F-F	4.9	2.3	Pop music. All instruments in stereo channels, heavy bass in centre channel, backing vocals and percussion in surround channels.
Radio drama	The Hitchhikers Guide to the Galaxy	F-F	1.1	1.8	BBC radio documentary. Speech and effects in stereo channels, effects in centre and surround channels. Lots of panning in surround channels.
Classical music	Last Night of the Proms	F-B	3.2	3.6	BBC live classical music recording. Full orchestral section in frontal channels. Reverberation in surround channels.
Film	Tropic Thunder	F-F	8.5	5.3	Action film. Effects and music in stereo channels, dialogue and effects in centre channel, effects in surround channels.
Ambient sound*	Thunderstorm	F-F	0.4	3.0	From "A Surround Sound Experience". Rain and thunder in stereo channels, rain in surround channels.
Pop music*	The Doobie Brothers - Long Train Runnin'	F-F	4.8	2.8	Pop music. All instruments in stereo channels, Guitar and drums in centre channel, guitar and vocals in surround channels.

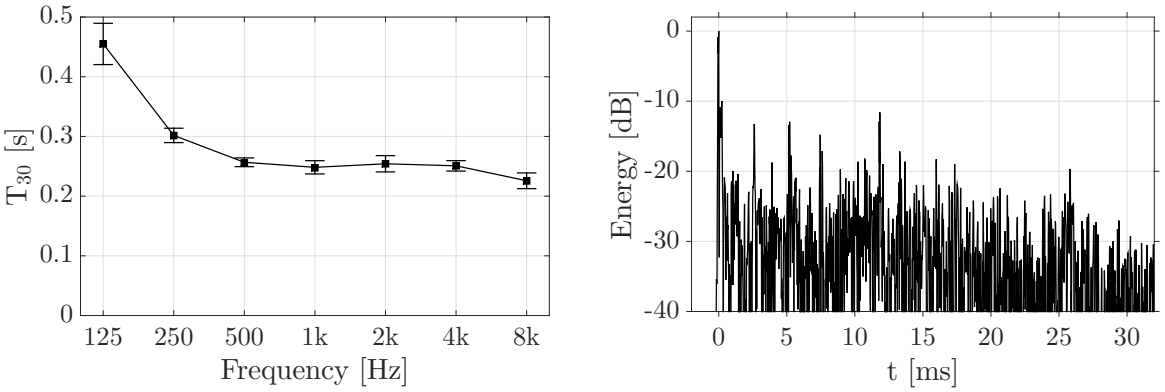
system. All stimuli were initially aligned to −23 LUFS (EBU, 2014) although were altered in the calibration procedure described in Section 4.3.6. In order for the 5.0 material to be used with the two soundbars and to be transmitted via digital optical cables, it was necessary to encode the stimuli. The stimuli were encoded to DTS digital surround with a SurCode DTS software encoder. The encoded 5.0 DTS WAV files had a sample rate of 44.1 kHz and a data rate of 1.234 Mb/s. All four reproduction systems received encoded files that were encoded with the same settings.

4.3.3 *Participants*

A total of 18 participants (age range: 21-42, mean 28, gender: 13 male, 5 female) participated in this study. All participants were fluent in English and self-reported normal hearing. A distinction between experienced and naïve assessors was made. Experienced assessors were professionals or academics in the fields of audio, acoustics or music and had previously participated in at least one other critical listening test. Naïve assessors were those who did not meet this requirement.

4.3.4 *Room*

Due to the nature of the technology used by the soundbars, an ITU-R BS.1116 (ITU-R, 2015b) standardised listening room was deemed unsuitable. Such a room has minimal side wall reflections so could possibly prevent lateral reflections produced by the soundbars. Instead, a user testing lab was used which was initially designed to represent a typical living space. It's dimensions on the horizontal plane can be seen in Figure 4.5, with a height of 2.9 m. The floor of the room was carpeted, the front and rear walls (from the perspective of the participant) were made of glass, and the side walls were typical plaster walls, one of which had a large mirror. A heavy curtain was drawn across the front wall. To quantify the acoustic properties of the room, acoustic measurements were conducted. The reverberation time (T_{30}) can be seen in Figure 4.4a and the energy-time curve can be seen in Figure 4.4b. For the mean reverberation time, three source positions and four receiver positions at each source position were measured. For the energy-time curve, the centre discrete speaker was used as the source and the listening position was used for the receiver position.



(a) (b)

Figure 4.4: Acoustic measurements of the room used. a) Mean reverberation time (T_{30}) in octave bands. Error bars show 95% confidence intervals. b) Energy-time curve relative to direct sound. y-axis values correspond to $10\log_{10}(p(t)^2)$, where p is pressure and t is time.

4.3.5 *Setup*

The laptop used to administer the listening test software was connected to a Roland UA-25EX interface running at a sample rate of 44.1 kHz and a bit depth of 16-bit. The digital audio signal from the interface was routed to a Friend-chip DMX-12 optical matrix via an optical Toslink cable. This optical matrix enabled switching between the various reproduction systems and was controlled via a MIDI signal output from the listening test software and a second UA-25EX interface. The optical matrix was connected to the various reproduction systems via optical Toslink cables. For the discrete surround and stereo systems, a Denon DN-A7100 AV receiver was used to decode the audio files and route the channels to the corresponding loudspeakers.

A schematic of the setup can be seen in Figure 4.5. The discrete surround system was composed of five PMC DB1S-A loudspeakers and a PMC TLE1 subwoofer. It was setup with a radius of 2.5 m and with angles specified in ITU-R BS.775 (ITU-R, 2012a) (i.e. 0°, ±30°, ±110°). To minimise elevation differences between the various systems, the loudspeakers were placed upside down with the tweeters below the mid-range drivers. The tweeter height was 1.07 m which corresponded to approximately ear height. The two soundbars, the centre speaker for the discrete system and the subwoofer for soundbar 1 were all placed on a desk with dimensions 1.5 x 0.6 x 0.85 m (w x d x h). The soundbars and discrete centre speaker were stacked with soundbar 1 at 0.92 m from the floor, soundbar 2 at 1.01 m from the floor and the tweeter from

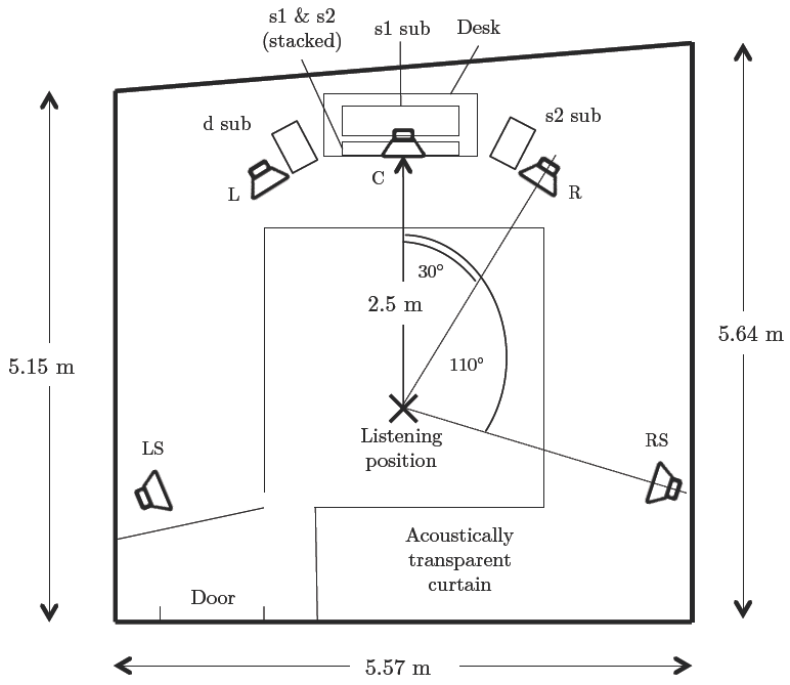


Figure 4.5: Room dimensions and experimental setup schematic.

the discrete centre channel at 1.07 m from the floor. There was therefore an elevation difference between the systems of 15 cm (~ 3° relative to the listener). All systems were 2.5 m from the listening position and an acoustically transparent curtain was used to prevent visual bias.

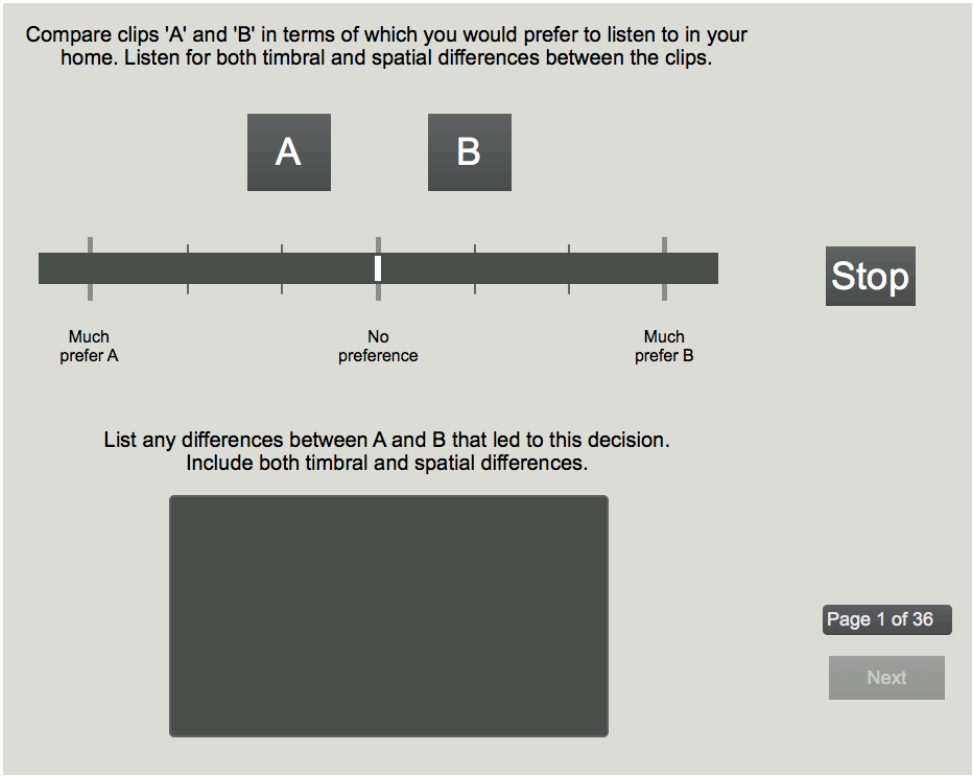
4.3.6 *Calibration*

The loudspeakers in the discrete system were level aligned relative to each other (to within 0.5 dB) using pink noise measured at the listening position with a sound pressure level meter. The subwoofer for the discrete system had a cutoff frequency of 85 Hz and was level aligned using pink noise and a 1/3-octave real time analyser to achieve a flat frequency response. The bass management was done via the AV receiver. A comfortable listening level was used for the absolute level of the discrete system (and therefore the other systems), as subjectively decided by the researcher and verified by participants in a pilot study. This corresponded to a mean stimuli level of 64.8 dBA for the discrete surround items. Soundbar 1 included several manual options to calibrate the soundbar for the room of use. These were set as - distance: 2 (medium), position: 6 (free standing), room: 8 (medium), subwoofer: 11 (dimension). There was

no automatic calibration process with this system. The level of the subwoofer was adjusted in the same manner as with the discrete system. Soundbar 2 was automatically calibrated using the inbuilt calibration function. This involved placing a small microphone (provided) at the listening position and running a short calibration program so that the soundbar could optimise the beam directions. Inter-system level alignment was carried out subjectively by the researcher. A subjective approach was chosen so as to take into consideration any psychoacoustic filtering employed by the soundbars. This was checked for every stimulus and gain was applied to the pre-encoded source files accordingly.

4.3.7 *Administration*

The experiment was administered via a laptop running a custom designed listening test patch using Max MSP software, see figures 4.6a and 4.6b. For both rating sessions each page contained two excerpts: "A" and "B", representing the same content for two different reproduction systems. The participants could play these as many times as they liked. It was possible to switch between the excerpts, although when a new excerpt was selected, it would always play from the beginning. The order of the comparisons was randomised, as was the assignment of stimuli to either button "A" or "B". The rating scale used for both sessions was based on a Comparison Category Rating scale (ITU-T, 1996) and had a range of values from -35 to $+35$ in interval increments. There were labelled anchors at -30 ("Much prefer A"), 0 ("No preference") and $+30$ ("Much prefer B") and unlabelled anchors at ± 10 and ± 20. The labelled anchors were offset from the end so as to reduce end-of-scale effects (Zielinski, Rumsey, and Bech, 2008). The scale had to be clicked before the participant could move on to the next page. Below the rating scale was a text box in which the participants could type any differences between the stimuli they heard. They were encouraged to list words or short phrases, not full sentences in order to reduce the test duration. The attribute rating interface had scales corresponding to the developed attributes of each participant. They were asked to "compare clips 'A' and 'B' in terms of which has more of the following attributes". The labelled anchors were "A has much more" and "B has much more". One rating scale had to be clicked before moving on to the next page.



(a) Preference rating and attribute elicitation interface.

(b) Attribute rating interface with example attributes.

Figure 4.6: Preference rating and attribute rating user interfaces.

4.3.8 *Data Collection*

Upon selecting 'Next' on each page of the interface, a participant specific time-stamped comma-separated values (CSV) file was appended with data from the previous page. For the preference rating pages this data consisted of page number, preference rating, stimulus 'A' name, stimulus 'B' name and text output from the text box. Additionally, at the end of the preference rating session a separate CSV file was created which contained all of the text data for refinement purposes. For the attribute rating pages this data consisted of page number, stimulus 'A' name, stimulus 'B' name, attribute ratings and corresponding attribute names. No further usage data was collected from the interface.

4.4 RESULTS

4.4.1 *Participant Reliability*

In paired comparison tests, it is possible to asses intra-participant reliability by using circular error rates (Parizet, 2002). A circular error occurs when a participant makes an inconsistent judgment on a triad of stimuli. For example, a circular error would occur if a participant preferred stimulus A to B, preferred stimulus B to C, but preferred stimulus C to A, i.e.

$$A \rightarrow B \rightarrow C \rightarrow A, \tag{2}$$

where $\rightarrow$ represents "is preferred to". Such errors indicate that either the participant was not paying attention, that they altered their assessment criteria as the test progressed, or that they found the test challenging resulting in inconsistent judgments. By comparing the number of circular errors associated with each participant and the maximum possible number of circular errors, a circular error rate in percent can be calculated. Figure 4.7 shows the circular error rates from the preference rating session for all participants. An error tolerance threshold of $A = 0.06$ was used for the calculations. If $A = 0$, no inverted preference can be accepted, whereas if $A > 0$, only preference inversions greater than the value of A are counted as circular errors. As a continuous slider was used in this study, a value of $A = 0.06$ was chosen (for scores scaled to ± 1) so as to discount slider inaccuracies of participants who intended to show no preference; $A = 0.06$ being equivalent to two interval increments either side of 0 on the scale. It is seen that the majority of participants could make reliable judg-

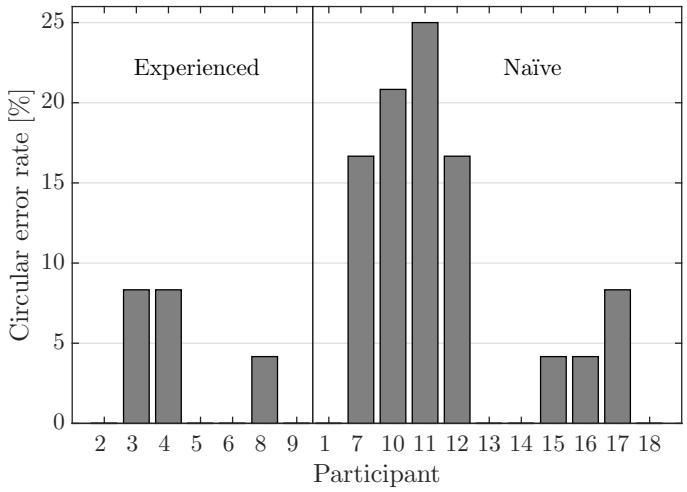


Figure 4.7: Circular error rates from the preference rating session calculated with an error tolerance threshold of $A = 0.06$.

ments with eight participants making no errors and 14 participants having an error rate of less than 10%. Participants 7, 10 11 and 12 (all naïve) show an error rate above 15%, which is higher than the average (mean) of 6.5%. To ensure reliability of results, it was decided to exclude the results from these participants for the subsequent analysis. An error limit of 10% was also used in (Woodcock, Moorhouse, and Waddington, 2014) to ensure consistent results from paired comparison ratings.

4.4.2 *Preference Ratings*

The preference ratings for each paired comparison were first scaled to lie in the range of ± 1, where -1 corresponds to full preference for stimulus "A" and $+1$ corresponds to full preference for stimulus "B". If P_{ij} is the preference probability of stimulus i versus stimulus j, it is assumed that

$$P_{ij} = -P_{ji}. \tag{3}$$

That is, a negative probability of preference P_{ij} means stimulus j is preferred to stimulus i. From these preference probabilities, preference scores can be calculated with

$$S_i = \sum_{j \neq i} P_{ij}, \tag{4}$$

where S_i is the preference score for stimulus i. As four reproduction systems were used, possible values of S_i lie in the range of ± 3. These preference scores were scaled to lie in the range of ± 1 so that $+1$ corresponds to full preference towards a reproduction system. In total, 24 preference scores were calculated (four systems by six content items).

After the exclusion of four participants, data from 14 participants remained (seven experienced, seven naïve) who rated four reproduction systems for six content items. To investigate the effect and interaction of the independent variables, a three-way mixed ANOVA was carried out on this data. The between-subject factor was participant experience (two levels) and the within-subject factors were system (four levels) and content (six levels). Prior to this, the main assumptions underlying the ANOVA were checked for each group of participants. Normality was checked using a Shapiro-Wilk test for the 24 combined within-subject factors for both groups. The data was not significantly different from normal ($p > .05$) for 22 out of the 24 factors for both the experienced and naïve groups. The data is therefore said to have a predominantly normal distribution. Homogeneity of variance was checked via Levene's test and the data was found to have equal variance ($p > .05$) for both within-subject and between-subject factors. Finally, Mauchly's test of sphericity was conducted which gave non-significant ($p > .05$) results.

The between-subject factor of participant experience was found to have a non-significant influence on the preference scores [$F(1, 12) = 2.68, p = .127$]. However, a partial eta-squared value of $\eta_p^2 = .183$ indicates that participant experience does have an effect on the preference scores. This suggests that if a larger study was conducted with more participants per group, participant experience could prove to be a significant factor.

An overview of the within-subject factor results from the mixed ANOVA model are shown in Table 4.2. For a significance level of 0.05, it is seen that factors System and Content*System are significant with values of [$F(3, 36) = 84.323, p < .001$] and [$F(15, 180) = 5.557, p < .001$] respectively. This shows that both the reproduction system and the interaction between the reproduction system and content have a significant influence on the preference ratings. All other factors and interactions are non-significant ($p > .05$). The fact that the interaction System*Experience is non-significant [$F(3, 36) = 2.403, p = .0084$] shows that in this study both groups of participants, experienced and naïve, rated the reproduction systems in similar ways. However, with

Source		df	F	η_p^2	p
System	Hypothesis	3	84.323	.875	$< .001$
	Error	36			
Content*System	Hypothesis	15	5.557	.317	$< .001$
	Error	180			
Content*System*Experience	Hypothesis	15	1.681	.123	0.058
	Error	180			
Content	Hypothesis	5	2.101	.149	0.078
	Error	60			
System*Experience	Hypothesis	3	2.403	.167	0.084
	Error	36			
Content*Experience	Hypothesis	5	0.303	.025	0.909
	Error	60			

Table 4.2: Within-subject factor results from the mixed ANOVA model in order of decreasing significance.

a partial eta-squared value of $\eta_p^2 = .167$, it could be the case that in a larger study a significant interaction could be found.

Post hoc Bonferroni-corrected pairwise comparisons of the significant factors System and Content*System were calculated. Figure 4.8 shows the marginal means for both the systems averaged over content and for each content individually. These preference scores, which can have values in the range of ± 1, are averaged over listener. Additionally, matrices that show which pairwise comparisons are statistically significant are presented. For the content average plot, all systems have corresponding marginal means significantly different from each other; the discrete surround system is the most preferred followed by the stereo down-mix. As an interaction was found between the reproduction systems and content, it is necessary to examine the marginal means for each system with respect to content. It is seen that the discrete surround system is significantly preferred over the stereo down-mix for the items "Ambience", "Pop Music", "Film" and "Radio Drama". These items are all F-F items whereas the items which do not show a significant difference are F-B items. The discrete surround system is significantly preferred to both soundbars for all content. Likewise, the stereo down-mix is significantly preferred to soundbar 1 for all content. For comparisons between the stereo down-mix and soundbar 2, only items "Pop Music" and "Radio Drama" show significant differences - the stereo down-mix being preferred in both cases. The only content which displays a significant difference between the two soundbars is the item "Ambience" for which soundbar 2 is preferred over soundbar 1. It should be noted

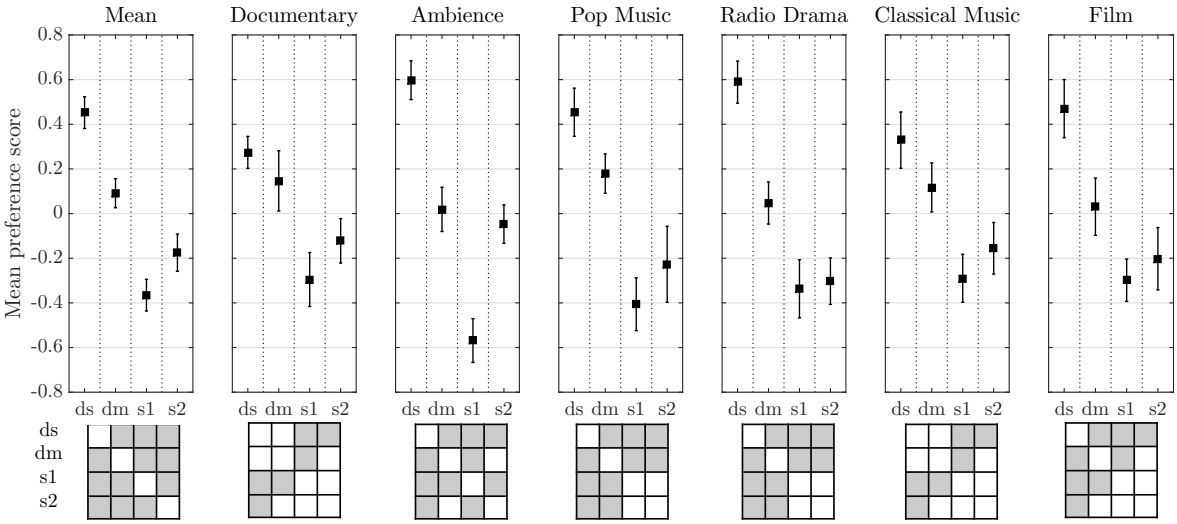


Figure 4.8: Marginal mean scaled preference scores for each reproduction system with respect to content, averaged over participant, with significance matrices. ds = discrete surround, dm = stereo down-mix, s1/2 = soundbar 1/2. Error bars show 95% confidence intervals. Shaded cells indicate pairwise significance.

that the greatest difference in preference scores is seen for the content item with the lowest frontal-surround RMS ratio (content "Ambience"), i.e. the content item with the most surround information.

4.4.3 *Sensory Profiling*

In the attribute elicitation and refinement stage the 14 participants whose preference ratings were analysed in the previous section produced a total of 102 refined attributes (max eight, min five). A visualisation of these refined attributes is presented in Figure 4.9. The raw paired attribute ratings were converted to attribute scores using the same method as with the preference scores. For each participant this resulted in an M x N matrix (or configuration) of attribute ratings, where M is the number of test items (24 in this case) and N is the number of individual attributes. The individual participant matrices were concatenated to form a complete attribute matrix of 24 items x 102 attributes. The following analysis was run on this dataset and implemented in XLSTAT.

The first stage of the analysis was to run Generalised Procrustes Analysis (GPA) on the dataset of attribute ratings. The aim of GPA is to reduce scale effects and to obtain a consensus configuration. This is achieved by rotating and transforming the configurations by minimising the residual distance between the configurations and

Figure 4.9: Word cloud of refined elicited attributes. The more frequently an attribute appeared, the greater prominence it has in the word cloud.

Table 4.3: Variability of eight principal components needed to describe 100% variance in the GPA model (3 sig. figs.).

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Variability (%)	75.7	12.3	3.68	2.75	2.13	1.72	1.08	.665
Cumulative (%)	75.7	88.0	91.7	94.4	96.5	98.3	99.3	100

their consensus. The second part of the analysis was to run Principal Component Analysis (PCA) on the dataset from the GPA procedure.

By identifying the "elbow point" in the cumulative variance data from the PCA analysis, it can be decided which components should be used to form the perceptual space. Components that appear before the elbow are retained for further analysis (Lawless and Heymann, 2010). A total of eight components were needed to explain 100% variance in the GPA model, Table 4.3. The first two components describe the majority of the variance, 88%, so are used to form the perceptual space.

Figure 4.10 shows all the rated attributes in the perceptual space of PC1 and PC2. The further the attributes are from the centre, the greater their associated explained variance. The inner and outer circles represent 50% and 100% explained variance respectively. Several clusters of attributes can be identified and are labelled with relevant descriptions. It is seen that PC1 (75.71% explained variance) is positively loaded with attributes related to width ("width", "wide", "width of frontal image", "width of image") and negatively loaded with the attribute "focussed". A cluster related to

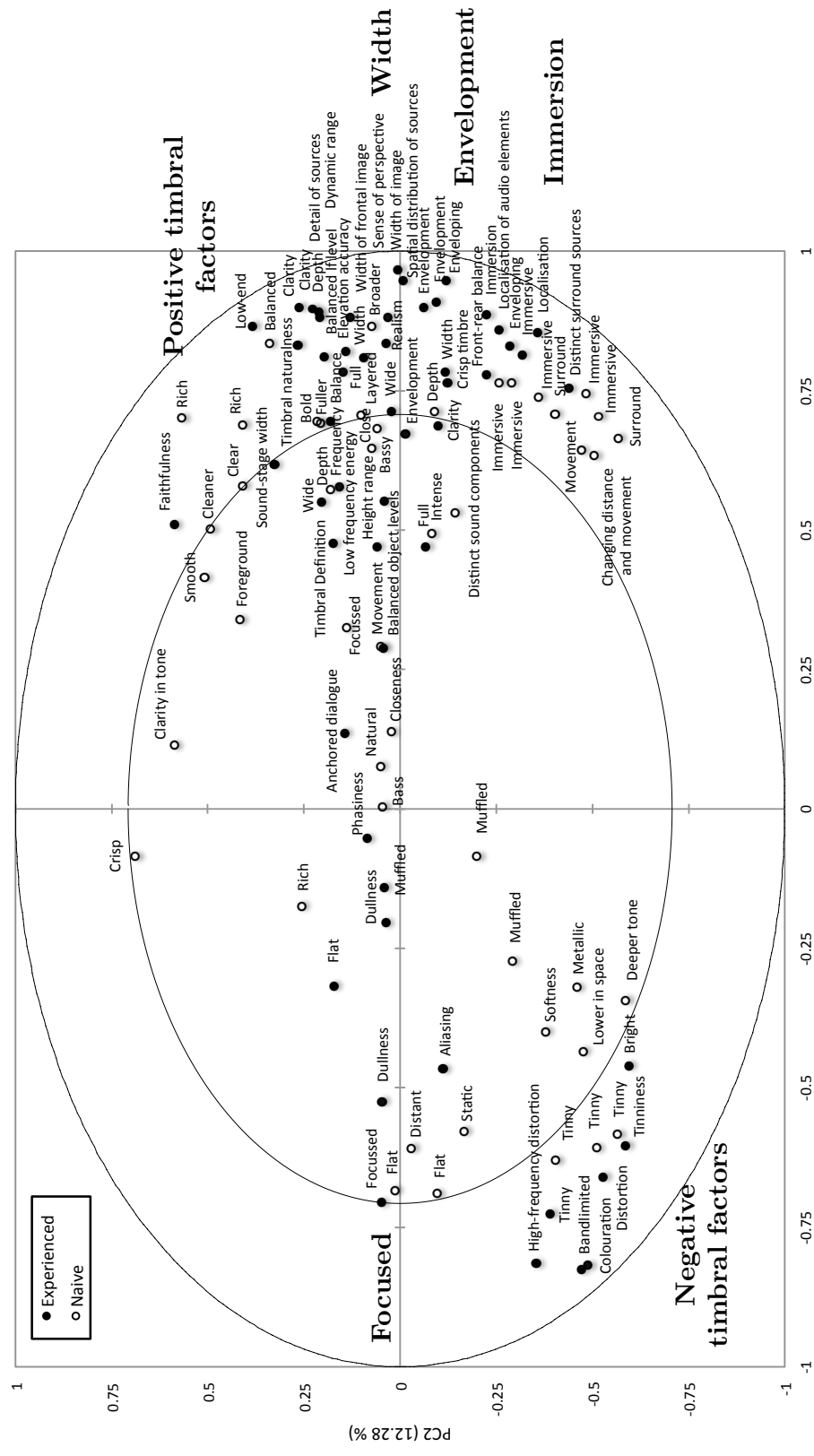


Figure 4.10: PCA correlation loadings with attributes in the space of PC1 and PC2. The inner and outer circles represent 50% and 100% explained variance respectively.

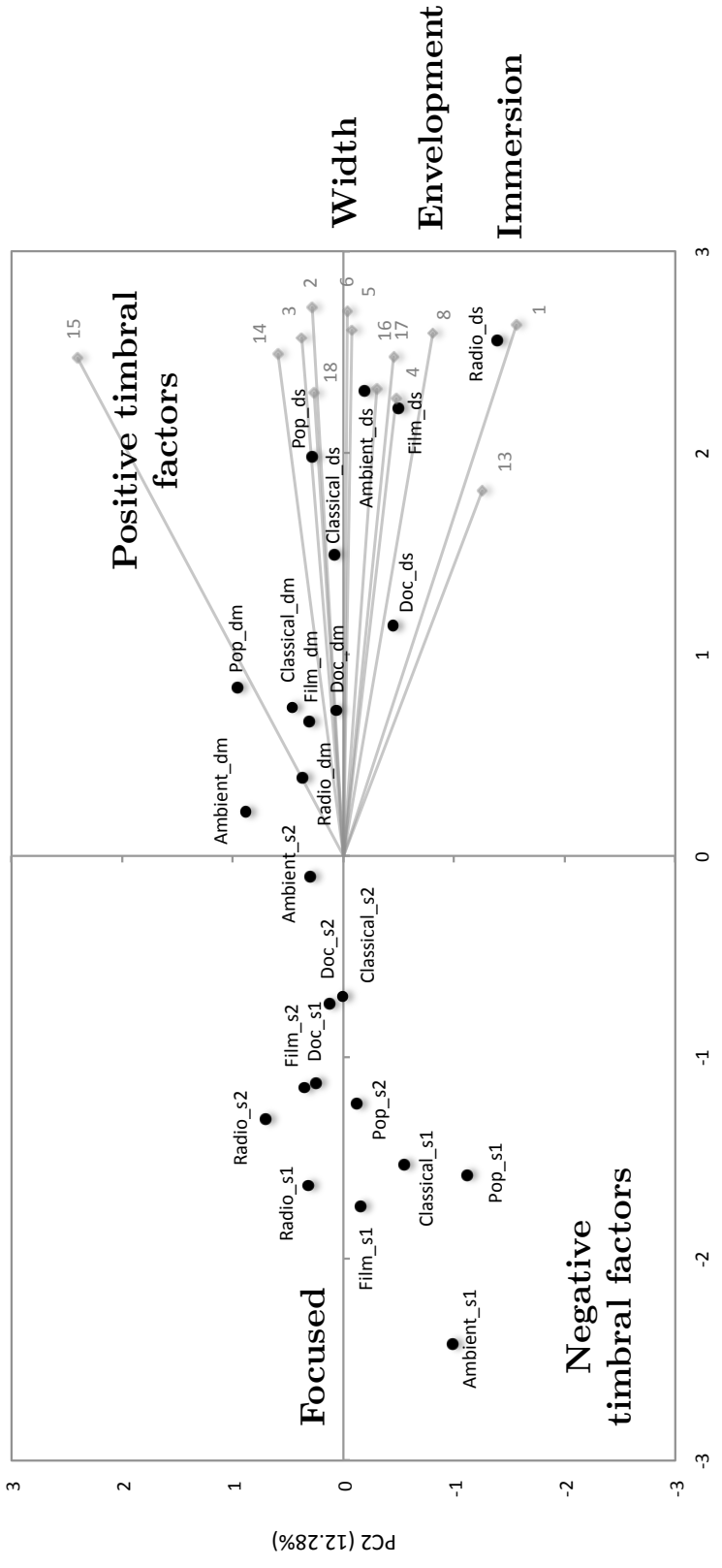


Figure 4.11: Objects and participants' preferences in the preference map of PC1 and PC2. ds = discrete surround, dm = stereo down-mix, s1/2 = soundbar 1/2.

envelopment is found in very close proximity to the width cluster ("envelopment", "enveloping") and a cluster related to immersion is also found close by ("immersion", "immersive", "localisation of audio elements", "changing distance and movement", "movement"). The grouping of these attributes suggests that envelopment terms were used in a similar way as width terms. Immersion terms however, were used slightly differently with attributes related to movement and localisation of audio elements playing a role. This distinction between immersion and envelopment was also commented on by several of the participants during the attribute elicitation stage. Clusters related to timbral aspects are found on the diagonals of the perceptual space. Negative timbral aspects ("bandlimited", "tinny", "colouration", "distortion") are found at negative PC1 and PC2 values whereas positive timbral factors ("rich", "clarity", "depth", "balanced") are found at positive PC1 and PC2 values. The fact that spatial and timbral aspects of the same polarity are located on the same sides of the perceptual space suggests that stimuli related to positive spatial attributes are also related to positive timbral attributes and *vice versa*. When comparing the attributes generated by experienced listeners (55) to those generated by naïve listeners (47), it is seen that similar terms are used, although the attributes from experienced listeners generally describe more variance in the data. This suggests that experienced listeners produce stronger attributes and are more confident in their use.

Figure 4.11 shows all the stimuli in the perceptual space of PC1 and PC2 with participants' preferences mapped through external preference mapping. The attribute cluster labels identified from Figure 4.10 are also shown. It is seen that items are generally clustered by the reproduction system. The discrete surround system stimuli are located in the area of the perceptual space related to width, envelopment and immersion. Stimuli such as "Classical_ds" and "Ambient_ds" are perceived as wide and enveloping whereas "Radio_ds" is the most immersive stimulus. At negative PC1 values soundbar 1 stimuli are found. These stimuli are also the most related to negative timbral factors. This suggests that soundbar 1 was not perceived as immersive or enveloping and had negative timbral characteristics. In particular, "Ambient_s1" is seen to be the most focussed and worst rated in terms of timbral aspects. This agrees with discussions from several participants. Soundbar 2 stimuli are also found at negative PC1 values suggesting that they are also not as wide, enveloping or immersive as the discrete surround system stimuli. "Ambient_s2", however, is found at a similar PC1 value to "Ambient_dm" which suggests that they are perceived as having similar widths. The fact that soundbar 2 stimuli have higher PC2 values suggests that they

are less related to negative timbral factors as soundbar 1 stimuli. In terms of spatial factors, soundbar 2 is seen to be perceived as slightly wider than soundbar 1. Finally, the stereo down-mix stimuli are perceived as less wide, enveloping and immersive than the discrete surround stimuli.

The participants' preferences mapped onto Figure 4.11 show that the majority prefer wide, enveloping and immersive stimuli (i.e. the discrete surround system stimuli). Participant 15, however, appears to have based their preference judgments more on timbral aspects than spatial aspects. This therefore means that their preference towards the discrete surround stimuli are less pronounced than with the other participants.

As the spatial and timbral factors shown in the perceptual space are not orthogonal, it is hard to separate their influence on the preference ratings of the various reproduction systems. It appears that the discrete surround system is preferred over the stereo down-mix due to spatial aspects, and this is expected as both systems use the same loudspeakers. The two soundbars are rated negatively compared to the other two systems due to a combination of both timbral and spatial factors. In terms of spatial factors, soundbar 2 is perceived as slightly wider than soundbar 1, although the only significantly different preference rating between these two systems is for the content "Ambience".

4.5 DISCUSSION

A subjective comparison of a discrete five-channel surround system, a discrete stereo system and two soundbars was made by means of a mixed methods approach. When averaged over content, preference ratings for the two soundbars were significantly lower than the discrete surround system and the discrete stereo system. Additionally, a significant difference in preference ratings between the two soundbars was observed; the soundbar which employs beamforming to achieve a surround effect was significantly preferred to the soundbar which predominantly uses filtering. With respect to content, the greatest difference between the systems was observed for the content item with the lowest frontal-surround ratio, i.e. the content item with the most surround information. This, along with the qualitative analysis, suggests that the two soundbars under study did not effectively replicate discrete surround channels.

Qualitative analysis showed that the given preference ratings were due to a combination of both timbral and spatial factors. Participants' preferences were mapped

to wide, enveloping and immersive items which correlated to the discrete surround system. It was not possible to fully separate the influence of spatial and timbral factors on the preference ratings as these clusters were not orthogonal on the perceptual space. This suggests that the systems with positive spatial attributes also had positive timbral attributes and *vice versa*.

When comparing results from the naïve and experienced listener groups, no significant difference in preference results was found. However, this may be due to the small sample size of seven participants per group and, as such, further larger scale experiments would be needed to explore this in greater detail. With the qualitative sensory profiling results, experienced participants produced on average approximately one more refined attribute (7.9) than naïve participants (6.7) and the attributes from experienced participants were seen to describe more variance in the data. This suggests that experienced listeners produced stronger attributes and were more confident in their use. Despite these differences, the semantics of the attributes produced by the two groups were seen to be similar.

Care must be taken when generalising the above results to the two types of technology as a whole. For a more comprehensive comparison a greater number of discrete systems and soundbars should be investigated to see if results are consistent with those found in this study. Additionally, It is thought that the room of use could have a significant influence on the effectiveness of soundbar technology. This should also be investigated in further studies.

In terms of the methodology, the modified OPQ method employed for this study was shown to be an effective method for investigating subjective differences between reproduction systems. Participants with a range of listening experience were able to give preference ratings, develop and refine individual attributes and rate stimuli on these attributes over two sessions. The qualitative data collected allowed an in-depth analysis of the reasoning behind the preference ratings.

Several modifications to the original implementation of the method were made, with the aim of making the method more suited to the application of comparing audio reproduction systems. Firstly, quality ratings were modified to preference ratings for the reason that when comparing audio reproduction systems without introduced degradations, it could be the case that quality is perceived as equally high for all systems, even though listeners may have certain preferences. Secondly, a paired comparison approach was taken throughout the modified method in comparison to a single stimulus approach in the original implementation. One advantage of using a paired

comparison method is that such an approach is able to provide a high degree of discrimination between stimuli whilst still being suitable for naïve participants. A high degree of discrimination is an important feature when comparing audio reproduction systems with participants who are not necessarily used to critical listening. Additionally, with a paired comparison approach it is possible to make consistency checks by means of circular error rates. Finally, the structure of the method was modified so that the preference rating and attribute elicitation stages occurred simultaneously. This meant that participants were only required to listen to the stimuli in two sessions rather than three, possibly reducing listener fatigue.

It is also worth discussing limitations of the modifications made above and the OPQ method in general. The number of comparisons to be made in a full factorial paired comparison approach rapidly increase with the number of systems to be assessed. This puts a lower limit on the number of systems than can be assessed with a paired comparison approach compared to a single stimulus approach. A limitation of using naïve participants for sensory profiling is that they are less acute to certain attributes compared to trained listeners. Depending on the purpose of the study, this may or may not be an issue. An alternative approach would be to use naïve participants for the preference rating stage and trained listeners for the sensory profiling stage.

As this method was shown to be an effective method for investigating subjective differences between reproduction systems, the method could be used for further studies on next generation audio perception. For instance, this method would be suitable to compare a range of other next generation audio reproduction systems, such as comparing different multichannel loudspeaker setups, comparing different next generation headphone techniques (e.g. binaural processing techniques) and also comparing conventional multichannel setups with more innovative reproduction methods such as media device orchestration. The latter of these applications would be particularly suited to this method of evaluation, as with soundbar versus discrete devices, there is the potential for two very different listening experiences. Conventional methods that rate "quality" and utilise provided attributes would therefore be limited in the insight they give about the listening experience, compared to the method used here. Rendering techniques are another device-related system IF that could potentially be compared with OPQ. Furthermore, a range of other system IFs could also be studied with this method. System IFs related to content such as the capture of live content by means of different microphone arrays could also be studied. Finally, it may be possible to investigate media-related system IFs, such as encoding and decoding, with this

method as the paired comparison approach allows for high discrimination. However, when assessing such small differences in quality, which could also be the case for some of the other applications mentioned above, the attribute elicitation stage could pose a challenge for naïve participants. As discussed above, this may mean the method would need to be adapted to account for this. For the other applications mentioned with more noticeable differences between conditions, the method should be suitable as is.

4.6 SUMMARY

In this chapter, the role of system influence factors on the quality of experience of next generation audio were explored. This was achieved by considering a comparison of soundbars - a next generation audio reproduction technology - and traditional discrete reproduction methods. With regards to this specific objective, it was seen that for the reproduction systems used the soundbars were less preferred than the discrete surround and stereo systems due to a combination of timbral and spatial factors. In addition, a suitable method by which to investigate the influence of system influence factors was applied from the field of multimedia QoE evaluation. This method was seen to be valuable for investigating subjective differences between reproduction systems and could therefore be used in further such studies, as well as to investigate the role of other system IFs on the QoE of next generation audio.

Part II

CONTEXT

5

ENVIRONMENTAL NOISE AND BACKGROUND-FOREGROUND AUDIO BALANCE

5.1 INTRODUCTION

In the next part of this thesis, we turn our attention to how next generation audio can be utilised to improve QoE in relation to context influence factors. As previously discussed, context IFs are factors that embrace any situational property to describe the user's environment and can be divided into physical, temporal, task, social, and technical and information factors. One physical context IF that is a key feature of many mobile listening environments is environmental noise. For this reason, to explore how next generation audio can be utilised to improve QoE in relation to context influence factors, in this chapter we investigate how object-based audio could allow for improvements in QoE when consuming audio in contexts with environmental noise present.

Mobile devices such as smartphones, tablets and laptops are playing an increasingly prominent role in the consumption of broadcast media with audio content. One consequence of this is that such media is being consumed in a wide range of contexts. As we saw earlier in this thesis, this trend towards ubiquitous listening is one aspect of next generation audio that may have large implications for its development. Whether it be in a café or on a train, the characteristics of the context of use are likely to influence the quality of experience for the consumer, as was discussed in Section 3.3.2. It was also seen in Section 2.2.3 that object-based audio offers the possibility to adapt audio content to better suit the system, environment and user. This, combined with the trend of extensive use of audio devices with significant processing capabilities, means that with object-based audio it may be possible to adapt an audio mix at the point of consumption so as to improve the listening experience for a given context. Prior to the research presented here, this concept had not been empirically studied, so in order to explore the role of context IFs on the QoE of next generation audio, it is this concept that is considered.

More specifically, the relationship between environmental noise and preferred background-foreground (BG-FG) audio object balance for headphone listening is investigated. This is relevant as environmental noise, i.e. extraneous noise associated with a given environment, is a key feature of many contexts where mobile listening occurs. Furthermore, adjusting BG-FG audio object balance is a very simple approach to adapting content that could be suitable for a range of content and devices. This research question is therefore related to two behavioural aspects of next generation audio that were identified earlier in this thesis - ubiquitous listening and the use of audio devices with significant processing capabilities. Additionally, this investigation illustrates how personalised and adaptive experiences, key experiential features of next generation audio, are associated with context influence factors.

The case study of investigating the benefits of object-based content adaptation with respect to environmental noise is a useful exemplar for studying context IFs in general for multiple reasons. Object-based content adaptation could be relevant for improving the listening experience with respect to a wide range of other context IFs. For instance, content could be adapted in various ways to react to social context IFs (e.g. persons present), task context IFs (e.g. multitasking), temporal context IFs (duration of use) and other physical context IFs (e.g. location). Object-based adaptation is thus a useful case study as there may be possibilities to extend the methods and approaches used for this particular application for investigations on other context IFs. Moreover, as environmental noise is a key characteristic of many listening situations it is likely that it is an important context IF for the QoE of other next generation audio technologies, for instance binaural techniques when in mobile situations. Again, methods and approaches used for this particular case study could therefore be beneficial for other such studies.

This chapter presents three empirical studies that investigate the above research question. The first of these is a laboratory-based listening test in which 22 participants were required to adjust the mix of audio content to their preference, whilst in the presence of reproduced environmental noise conditions. The aim of this was to establish whether environmental noise has a significant influence on preferred BG-FG balance. The second experiment is of a similar nature to the first; a laboratory-based listening test with 22 participants. In addition to the adjustment task (with different experimental conditions to the first experiment), semi-structured interviews were conducted to probe the reasoning behind participants' adjustments. Finally, the third experiment is a web-based listening test in which there were 50 participants. This experiment

consisted of both listening tasks and qualitative survey questions. By considering the quantitative and qualitative data spanning the three experiments, an understanding of both how and why users adjust audio mixes in noisy environments has been developed.

Before presenting these studies, relevant literature that has not previously been presented in this thesis, namely on the topics of audio adaptation for mobile listening and background/foreground object distinction, is discussed.

5.1.1 *Audio Adaptation for Mobile Listening*

Adapting audio to improve the listening experience from mobile devices is not a novel concept in itself and, indeed, a range of approaches addressing this topic have previously been presented. On the one hand, there are methods that aim to improve the audio quality from loudspeakers incorporated into mobile devices. Such methods include low frequency enhancement (Cecchi et al., 2016; Turnbull, Hughes, and Hoare, 2008), spatialisation (Cecchi et al., 2016; Drossos et al., 2012), dialogue clarity improvement (Czyzewski et al., 2016), dynamics processing (Czyzewski et al., 2016; Turnbull, Hughes, and Hoare, 2008) and linearisation of frequency response (Cecchi et al., 2016; Czyzewski et al., 2016; Turnbull, Hughes, and Hoare, 2008). These methods attempt to compensate for the constraints of the devices themselves and, with the exception of (Czyzewski et al., 2016), do not aim to account for the listening environment.

Other approaches to improve the listening experience from mobile devices focus on headphone listening and, more specifically, how to reduce the impact of environmental noise on the audio experience. It is these methods that are most relevant to this study. Active noise-cancelling headphones are perhaps the most well known technology to address this issue and are commonplace in the aviation industry (Molesworth, Burgess, and Kwon, 2013) as well as more recently appearing in consumer devices for recreational mobile listening. The complexity of noise-cancelling headphones, however, means that they are generally expensive and, furthermore, noise-cancellation may not be desirable in all mobile listening situations. A simpler approach can be found with volume-based and dynamic range-based methods, as previously discussed in Section 3.3.2. Such methods utilise inbuilt microphones on mobile devices to monitor the environmental noise level of the listening environment and adjust the volume or dynamic range of the audio content accordingly, for exam-

ple see (Reis, Carriço, and Duarte, 2009), (Mason et al., 2015) and (Kean, Johnson, and Sheffield, 2015).

The concept presented in this chapter is a fundamentally different one to those discussed in the previous paragraph. In the previous approaches the audio content is modified as a whole; the entirety of the content is processed with the same volume-based or dynamic range-based algorithms. This has the advantage of simplicity - no information is needed about the content being reproduced. A drawback of such methods, however, is that there is limited flexibility with regards to accounting for personal preference, listening context and the content. In the approach presented here, it is the audio mix that is modified at the point of consumption to account for environmental noise conditions. Such an approach requires more metadata from the audio content, but in return a higher degree of flexibility can be achieved. It is the object-based audio paradigm that allows for such an approach.

5.1.2 *Background/Foreground Object Distinction*

It is clear that object-based audio allows for a wide range of content adaptations to be made at the receiver end. One of the simplest of these is to adjust the relative levels (balance) of sounds in a mix. Grouping audio objects into categories simplifies this process further and several categorisations of audio objects are found in the literature. In the context of spatial audio evaluation, Rumsey et al., (2008) distinguish between background components consisting of diffuse or environment-related aspects of the scene, and foreground components consisting of localisable objects. In the context of television audio for hearing impaired users, Shirley and Oldfield, (2015) propose three categories of audio objects - speech content whose comprehension is critical, background noise that has been shown to be detrimental to both clarity and to perceived overall sound quality, and other non-speech sounds that are considered important to comprehension and/or enjoyment of the material. In a more complex categorisation of broadcast audio objects, Woodcock et al., (2016) used hierarchical agglomerative clustering to identify seven general categories, which relate to sounds indicating actions and movement, continuous and transient background sound, clear speech, non-diegetic music and effects, sounds indicating the presence of people, and prominent attention grabbing transient sounds. In the studies presented in this chapter a simple background/foreground categorisation is used; foreground objects are important to

the narrative and generally localisable whereas background objects are non-critical to the narrative and generally more diffuse.

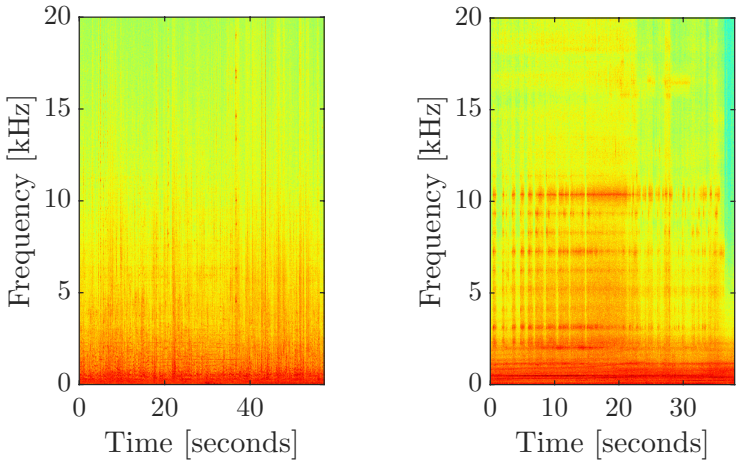
5.2 EXPERIMENTAL DESIGN: STUDY I

The aim of the first experiment was to investigate whether environmental noise has a significant influence on preferred BG-FG balance. A laboratory-based study was conducted in which participants adjusted the BG-FG balance of audio content to their preference whilst environmental noise was reproduced via a 3D loudspeaker setup. Participants made BG-FG adjustments for four environmental noise conditions and three audio excerpts, which were reproduced via two different methods. The following sections describe these conditions in more detail before the test procedure is outlined.

5.2.1 *Environmental Noise*

So as to have full control of the environmental noise, this study was carried out in a laboratory setting. This therefore posed the challenge of realistically reproducing environmental noise in a listening room. A range of studies from the field of soundscapes have investigated how to reproduce environmental noise in an ecologically valid way. Guastavino et al., (2005) state that it is necessary to provide both a neutral visual environment and a good sense of spatial immersion in order to ensure ecological validity when reproducing urban environmental noise. One method of providing a good sense of spatial immersion is through ambisonics - a reproduction method that works by capturing a loudspeaker-independent 3D representation of a soundfield and decoding this to a loudspeaker array. Studies have shown (Davies, Bruce, and Murphy, 2014; Guastavino et al., 2005) that ambisonics can successfully reproduce soundscapes in the sense of semantic evaluation, that is, similar semantic categories are elicited by reproduced soundscapes as *in-situ* soundscapes. Therefore in order to reproduce realistic environmental noise in this study, ambisonic reproduction was used.

In order for the results to have ecological validity, it was important to choose environmental noise clips that correspond to realistic use cases for mobile audio listening. With that in mind, two scenarios were chosen: a café type environment and an underground train environment - both situations where mobile listening is common. Table 5.1 outlines the properties associated with the noise clips. As well as representing dif-



(a) Café. (b) Train.

Figure 5.1: Environmental noise spectrograms.

ferent use cases, the two clips were chosen to be spectrally different, as seen in Figure 5.1; the train environment has more high frequency content than the café environment and is seen to be more tonal. The clips were trimmed so as to be spectrally consistent throughout their duration.

These two environmental noise clips were used to create a total of four noise conditions: "No noise", "Café quiet", "Café loud" and "Train". The "Café quiet" clip was calibrated to an L_{Aeq} of 54.5 dBA, the "Café loud" clip to an L_{Aeq} of 64.0 dBA and the "Train" clip to an L_{Aeq} of 64.8 dBA. The calibration levels of the two café clips were chosen to be representative of realistic levels. The calibration level of the "Train" clip was chosen to equal the level of the "Café loud" clip, although it should be recognised that this is possibly lower in level than real-world situations (Neitzel et al., 2009). The recordings were B-format and were provided by a professional sound recordist who captured them with a Soundfield ST450 microphone system. The four-channel B-format files were decoded for an eight-channel cube array with a WigWare Ambisonic Decoder (Wiggins, 2010).

5.2.2 *Audio Excerpts*

Three pieces of audio content were used in this study, as described in Table 5.1. Items "Sport" and "Radio Doc" were sourced from object-based productions, meaning that mixes of separate background and foreground objects were available. Item "TV Doc" was sourced from a 5.1 channel-based production. The centre channel provided the

Stimulus	Description	Background	Foreground
Café Noise	B-format recording made in a New York diner. Sounds of many conversations, distant music and occasional cutlery clatter.	-	-
Train Noise	B-format recording made on an underground train. Sounds of rumbling carriage, electric engine, screeching wheels and very distant platform announcements.	-	-
Sport	20 second excerpt of an English football broadcast	Crowd noises	Commentary
Radio Doc	15 second excerpt of a radio documentary - "The Cornish Gardner"	Music and atmospheres	Narration
TV Doc	17 second excerpt of a TV nature documentary - "Africa"	Orchestral music and effects	Narration and prominent effects

Table 5.1: Descriptions of environmental noises and audio items.

foreground content for this item and the remaining channels provided the background content. This was a suitable distinction as the dialogue in this clip was mixed to the centre channel. The LFE channel was discarded. All items were available in a five-channel surround format and were down-mixed to stereo via coefficients specified in ITU-R BS.775, Annex 4 (ITU-R, 2012a). The excerpts ranged in duration from 15 to 20 seconds so as to be long enough for judgments to be made but short enough for the constituent background and foreground levels to remain relatively constant. The excerpts were normalized to -23 LUFS (background and foreground combined) and all audio used was 24 bit, 48 kHz.

5.2.3 *Reproduction Methods*

The audio content was reproduced via two headphone-based methods: conventional two-channel stereo and virtual surround sound by means of dynamic binaural processing. Virtual surround sound aims to emulate a surround speaker setup over headphones by using head-related impulse responses (HRIRs) (McKeeg and McGrath, 1997). For increased plausibility, impulse responses containing the room response, known as binaural room impulse responses (BRIRs), can be used in combination with head-tracking, which keeps the reproduced scene stable by compensating for head motion (Lindau and Weinzierl, 2012). By comparing these two methods it will be pos-

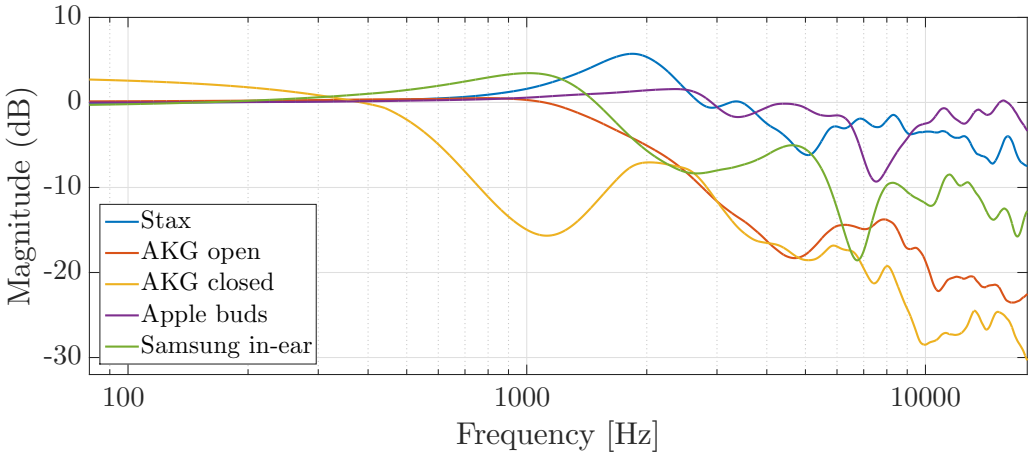
sible to assess if preferred BG-FG balance is dependent upon headphone reproduction method.

5.2.4 *Setup*

The experiment was carried out in an ITU-R BS.1116 compliant listening room (ITU-R, 2015b) with further properties outlined in (Nixon, Bonney, and Melchior, 2015). For the environmental noise reproduction a setup of eight Genelec 8030B loudspeakers was used, with four placed on the floor at the corners of a square with a 3.3 m side length and four placed on the ceiling at the corners of a square with a 2 m side length. The listening position was located at the centre of the array, both horizontally and vertically. Although the loudspeakers were not setup to form a perfect cube due to practicalities, this was not believed to be detrimental to the reproduction. Loudspeaker magnitude responses were equalised, time-aligned and level-aligned according to (ITU-R, 2015b). An acoustically transparent curtain was placed in a square around the listener so as to prevent visual bias and increase plausibility of the reproduced scene.

Open-back electrostatic headphones (STAX SR-207) were used. It should be noted that headphones of this brand and design are close to acoustically transparent (Satongar et al., 2015). In measurements made for this study, this was confirmed and it was found that average attenuation levels were similar to those of popular earbud-style headphones, see Figure 5.2. To obtain the attenuation responses shown in Figure 5.2, a Neumann KU100 dummy head microphone was used to record decorrelated white noise reproduced via the eight-channel speaker array, with and without headphones. For each set of headphones, measurements were made at three positions in the listening room. Attenuation functions were calculated by subtracting fast Fourier transforms (FFTs) of the measurements without headphones from FFTs of the measurements with headphones and averaging over the three positions. These functions were then smoothed with a Lowess model.

A BBC R&D binaural renderer was used in combination with an optical head-tracking system (VICON Bonita) to dynamically render the virtual surround sound content. The five virtual surround loudspeakers were positioned according to ITU-R BS.775 (ITU-R, 2012a). Binaural room impulse responses corresponding to the room and speaker setup in question were used in the binaural rendering. For additional



(a) Smoothed attenuation functions.

(b) Stax. (c) AKG open. (d) AKG closed. (e) Apple bud. (f) Samsung in-ear.

Figure 5.2: A comparison of attenuation functions (a) from various models of headphones (b-f).

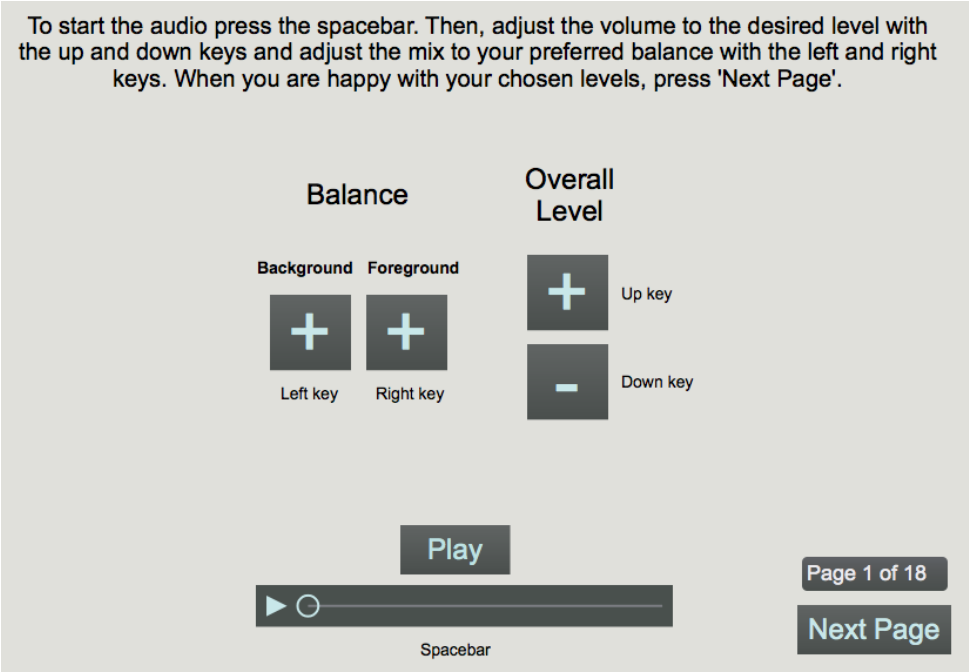


Figure 5.3: Graphical user interface.

information regarding the measurement of these binaural room impulse responses please refer to (Pike, Melchior, and Tew, 2014).

5.2.5 *Procedure*

The listening test software was implemented in MAX MSP, a screenshot of which can be seen in Figure 5.3. Each page of the software represents the adjustment of one audio excerpt for one environmental noise condition and one reproduction method. Controls available to the participant were BG-FG balance, overall level and play controls for the audio excerpt. The BG-FG balance was adjustable from only background objects audible to only foreground objects audible. Both the initial BG-FG balance and overall level were randomised. The playback of the environmental noise was not controllable by the participants. It should be noted that there was no visual feedback for the adjustments so as not to influence the participants' judgments.

A repeated measures design was used so that each participant had to make adjustments for all conditions. Two noise conditions were repeated so as to enable analysis of participant consistency, therefore a total of 36 adjustments were made per participant. The experiment was split into two sessions of 18 adjustments corresponding to each of the two reproduction methods, with a short break in between. The or-

der of reproduction method was balanced across participants. For each reproduction method, adjustments were grouped by the environmental noise condition, that is, the three audio excerpts were adjusted for each environmental noise condition in succession. When a new environmental noise began, participants had to wait for at least 20 seconds before playing the audio excerpts so as to familiarise themselves with the environmental noise. The order of the environmental noise and audio excerpt conditions were randomised. Before the main rating sessions, a familiarisation stage allowed participants to explore the interface and to listen to the three audio items without environmental noise.

5.2.6 *Data Collection*

Upon selecting 'Next page' on each page of the interface, a participant specific time-stamped CSV file was appended with data from the previous page. This data consisted of page number, background level, foreground level, FG-BG ratio, overall level, background stimulus name, foreground stimulus name and noise stimulus name, where all levels were expressed as decibels relative to full scale (dBFS). No further usage data was collected from the interface.

5.2.7 *Participants*

A total of 22 participants (age range: 19-45, mean: 28, gender: 11 male, 11 female) participated in the study. All participants self-reported normal hearing, were fluent in English and could be classed as naïve listeners, that is, they were not professionals in the field of audio and had no or little experience of critical listening tests. When asked about their mobile listening habits, all participants reported that they listen to audio content from a mobile device with headphones at least monthly, with 77% reporting that they listen to audio content from a mobile device with headphones everyday.

5.3 RESULTS: STUDY I

5.3.1 *Outlier Detection*

The recorded data consisted of background-foreground balance (referred to as FG-BG ratio in the results[1]) and overall level values for every condition. An initial analysis of the data revealed a number of extreme outliers (2.6% of all data), defined as values which lie outside three times the interquartile range (IQR). Due to the nature of these outliers it was suspected that the majority were due to an error in the test software. More specifically, the process that exported the adjustment data to text files occasionally did not initiate correctly resulting in 100% background or foreground adjustments being recorded. These occasional errors were confirmed by questioning participants about their responses and, as a result, it was deemed justifiable to remove these specific outlying data points from further analysis. As the initiation error resulted in 100% background or foreground adjustments being recorded, there was no concern about other data being unreliable due to this software error. Mild outliers were also observed (a further 2.4% of data), defined as values which lie outside 1.5 x IQR, although these were not excluded as they were considered as valid data.

5.3.2 *Participant Consistency*

Participant consistency was assessed by examining the mean variance between repeated adjustments. Noise conditions "No noise" and "Café loud" were repeated for all content and both reproduction methods resulting in a total of 12 repeated conditions per participant. When averaged over participants, the mean overall variance (including both ratio and level variances) was 3.9 dB. Participants were less consistent with ratio adjustments compared to level adjustments - the mean ratio variance was 4.5 dB compared to a mean level variance of 3.3 dB. In terms of individual participants, the most consistent participant had a mean overall variance of 2.1 dB and the least consistent had a mean overall variance of 5.9 dB. The distribution of participant consistency was assessed and from this it was decided to include all participants in the further analysis.

1 In (Walton et al., 2016) this is referred to as background-foreground ratio.

Variable	Source	df	F	p
Level	System	1/475.07	133.11	$< .001$
	Content	2/475.05	6.37	.002
	Noise	3/475.04	212.87	$< .001$
	System*Content	2/475.04	3.09	.046
Ratio	System	1/460.00	5.37	.021
	Content	2/459.30	34.78	$< .001$
	Noise	3/459.50	4.61	.003

Table 5.2: Statistically significant type III fixed effects for dependent variables level and ratio.

5.3.3 *Linear Mixed Model Analysis*

Due to the removal of outliers, a repeated measures analysis of variance could not be used to analyse the interaction and statistical significance of variables. Instead, a linear mixed model analysis was conducted for both the ratio and level data. As fixed effects in the model, variables System (reproduction method), Content (audio content) and Noise (environmental noise) including all interactions were used. To account for differences between individuals, variable Participant was used as a random effect in the model, including intercepts. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality.

Level

Firstly, the model was calculated with level as the dependent variable. Type III tests of fixed effects revealed that the main effects of System, Content and Noise were statistically significant ($p < .05$), as well as the interaction System*Content (see Table 5.2). In particular, we are interested in how the environmental noise influences the level and therefore *post hoc* Bonferroni-corrected pairwise comparisons of the significant factor Noise were calculated. From Figure 5.4 it is seen that, as expected, participants increased the level according to the environmental noise. This is consistent with other studies, e.g. (Breinbauer et al., 2012). The spectral characteristics of the noise did not have a significant influence on the level, as seen by the two noise items "Café loud" and "Train", which are not significantly different.

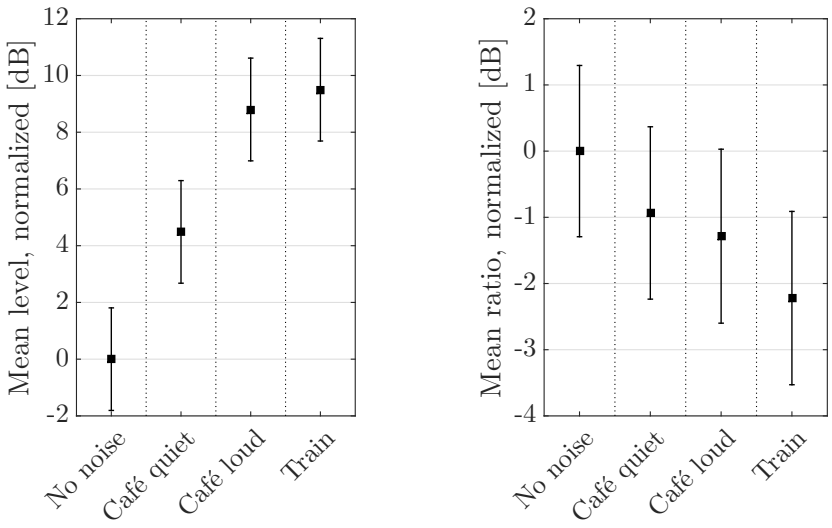


Figure 5.4: Mean level and ratio with respect to environmental noise, normalized to the mean of the "No noise" item. A positive ratio represents increased foreground levels whereas a negative ratio represents increased background levels. Error bars show 95% confidence intervals.

Ratio

For FG-BG ratio as the dependent variable the main effects of System, Content and Noise were all statistically significant (see Table 5.2). However, none of the interactions were. This therefore suggests that participants preferred different FG-BG ratios for the different noise conditions, the different systems and also the different pieces of audio content. To investigate how environmental noise influenced the preferred FG-BG ratio, the mean ratio with respect to environmental noise was examined, Figure 5.4. In other words, this is the mean ratio averaged over both content and system. *Post hoc* Bonferroni-corrected pairwise comparisons of the significant factor Noise reveal a significant difference between conditions "No noise" and "Train" ($p = 0.02$). Interestingly, the noise conditions have a mean ratio that is negative, which represents an increased background level in comparison to the "No noise" condition.

Despite the fact that the three-way interaction System*Content*Noise did not have a significant effect on the FG-BG ratio, it was considered useful to examine plots showing the ratio of each condition separately. To enable easier interpretation of the results, data from each participant were first normalized to the "No noise" conditions. In other words, for each participant the difference between the "No noise" condition and the other noise conditions for each piece of content and system was calculated. Figure 5.5

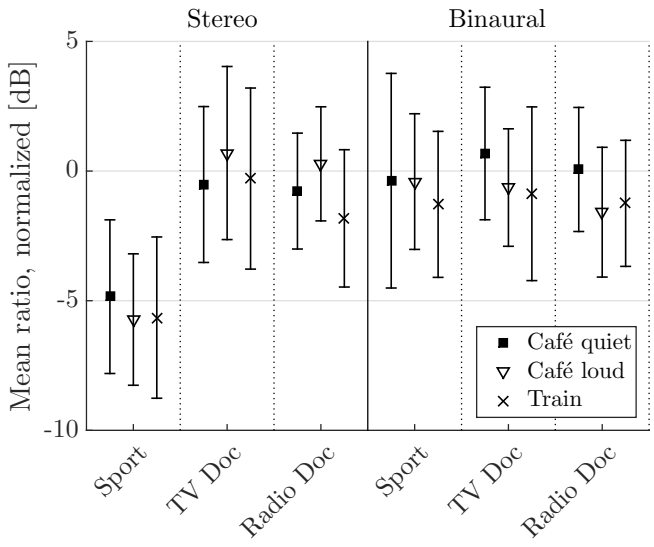
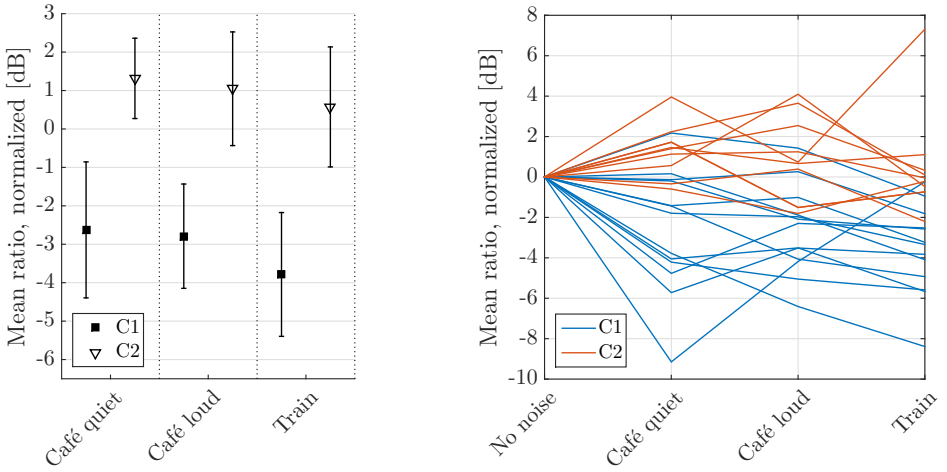


Figure 5.5: Mean ratio for all conditions, normalized to participants' "No noise" ratios. 0 dB represents the "No noise" ratio. Error bars show 95% confidence intervals.

shows the mean normalized ratio for each condition. It is apparent from Figure 5.5 that, even though no significant interactions were found in the linear mixed model calculated from the non-normalized data, interactions between System, Content and Noise should not be disregarded. Content "Sport" via stereo reproduction appears to exhibit different behaviour from the other Content and System combinations. Such interactions should be considered in further studies. It is also apparent from Figure 5.5 that there is a large degree of uncertainty in the results. From looking at data from individual participants and also from discussions with participants, it is believed that this large uncertainty is simply caused by differing preferences for the background-foreground ratio in noisy conditions; the majority of participants increased the background levels although others kept a constant ratio across all conditions or increased the foreground levels in noisy conditions. This was investigated further by applying a clustering algorithm to the ratio data.

5.3.4 *Cluster Analysis*

Cluster analysis is a statistical analysis method whereby objects are classified into clusters that share similar properties, see (Silzle et al., 2009) for an example in the context of audio preference tests. In order to investigate possible clustering in the ratio data, a k-means clustering algorithm was applied to the normalized ratio data as a whole with two clusters and a simple Euclidean distance measure. The two clusters



(a) Mean ratio for each cluster. Error bars show 95% confidence intervals.

(b) Mean ratio for each participant. Line colour represents cluster membership.

Figure 5.6: Mean ratio (normalized to the "No noise" item) with respect to environmental noise and cluster membership, averaged by cluster (a) and for each participant individually (b).

(C1 and C2) consisted of 13 and nine participants respectively. It should be noted that when clustering into three or four clusters, the majority of participants were still grouped into two clusters (clusters of 11, 8, 3 and 12, 7, 2, 1 participants respectively). Whereas this is not proof that the data does not consist of more than two meaningful clusters, it does suggest that analysis of more than two clusters would be limited due to the small group sizes. The ratio patterns of the two clusters can be observed in Figure 5.6. Cluster C1 consists of the same subset of participants for all noise types and likewise for cluster C2. It is seen that participants in cluster 1 adjusted the FG-BG ratio towards higher background levels in the presence of environmental noise, whereas participants in cluster 2 slightly increased the foreground levels or kept the ratio the same. It is apparent from the individual participant plot in Figure 5.6 that the two clusters should not be treated as completely distinct and isolated groups as participant responses from each cluster overlap. It is also apparent that the individual participant data is relatively noisy, meaning that participants do not always make consistent responses with regards to their cluster membership. Despite this, the cluster analysis suggests that FG-BG ratio preference is listener specific and analysis based on averages across all participants is not sufficient on its own.

5.3.5 *Discussion*

The results presented above have illustrated that environmental noise can significantly influence preferred BG-FG balance. It was shown that, when in the presence of environmental noise, the majority of participants adjusted the BG-FG balance towards higher background audio levels compared to adjustments made without environmental noise. This interesting result can possibly be explained from discussions that were had with the participants. One participant mentioned that they "used the crowd to drown it [the environmental noise] out", where the crowd refers to the background content in the "Sport" item. In other words, it appears that some participants were adjusting the balance so that the background audio masked the unwanted environmental noise, whilst keeping the foreground audio at an intelligible level. This raises the question of whether this trend reverses at higher environmental noise levels. Perhaps at moderate environmental noise levels, like those used in this experiment, intelligibility is not an issue and therefore the balance can be adjusted towards higher background levels to mask the environmental noise. At higher environmental noise levels however, it might be the case that intelligibility becomes more of an issue and therefore the balance is adjusted to higher foreground levels. This is investigated in study II, presented in the following section.

Another participant mentioned that they adjusted the balance towards more background audio so as not to miss any of the background objects that they felt were important to the mix. This raises questions about spectral similarities between the environmental noise and background audio objects. One reason why the "Sport" clip was often adjusted to much higher background audio levels with environmental noise could be due to the similarities between the crowd sounds (background) and the environmental noise. It was speculated that the dynamic range of the background audio might be higher than that of the foreground content and would therefore need to be increased more in the presence of noise. This was however checked and was found not to be the case.

5.4 EXPERIMENTAL DESIGN: STUDY II

Results from study I indicate that it may be possible to adapt object-based content in order to improve the listening experience in noisy environments. The large variance in results highlights the personal nature of the adjustments and, indeed, cluster analysis

indicated that participants may be grouped according to their preferences. The aim of the second study was to expand the results from study I with a focus on two points. Firstly, a greater range of environmental noise levels were used in order to investigate whether the trends seen in study I continue for higher noise levels. Secondly, a qualitative aspect was added in the form of semi-structured interviews in order to probe the clustering of participant responses.

The method used was similar to study I - a laboratory-based study in which participants adjusted the BG-FG balance of audio content to their preference whilst environmental noise was reproduced via a 3D loudspeaker setup. Participants made adjustments for nine environmental noise conditions and three audio excerpts. Unlike study I where two headphone reproduction methods were compared, only stereo headphone reproduction was considered.

5.4.1 *Environmental Noise*

As in study I, two types of environmental noise were used. The "Train" recording from study I was included, although the "Café" condition from study I was replaced with the condition "Crowd". This was due to the "Café" recording sounding unnatural at high sound pressure levels. The "Crowd" recording had similar spectral properties to the "Café" recording but was recorded in an environment with naturally higher sound pressure levels. These two recordings were reproduced at 65 dBA, 70 dBA, 75 dBA and 80 dBA. A "No noise" condition was also included. The recordings were ambisonic B-format and were decoded to an eight-channel cube loudspeaker array.

5.4.2 *Audio Excerpts*

The same three audio excerpts were used as in study I: "Sport", "Radio Doc" and "TV Doc". See Section 5.2.2 for details of these.

5.4.3 *Setup*

The headphones used in this study, AKG K702 open back headphones, differ to those used in study I. Whereas those used in study I were close to acoustically transparent in a similar nature to earbuds, these had attenuation properties of typical over-ear headphones. When comparing measured attenuation levels made with broadband

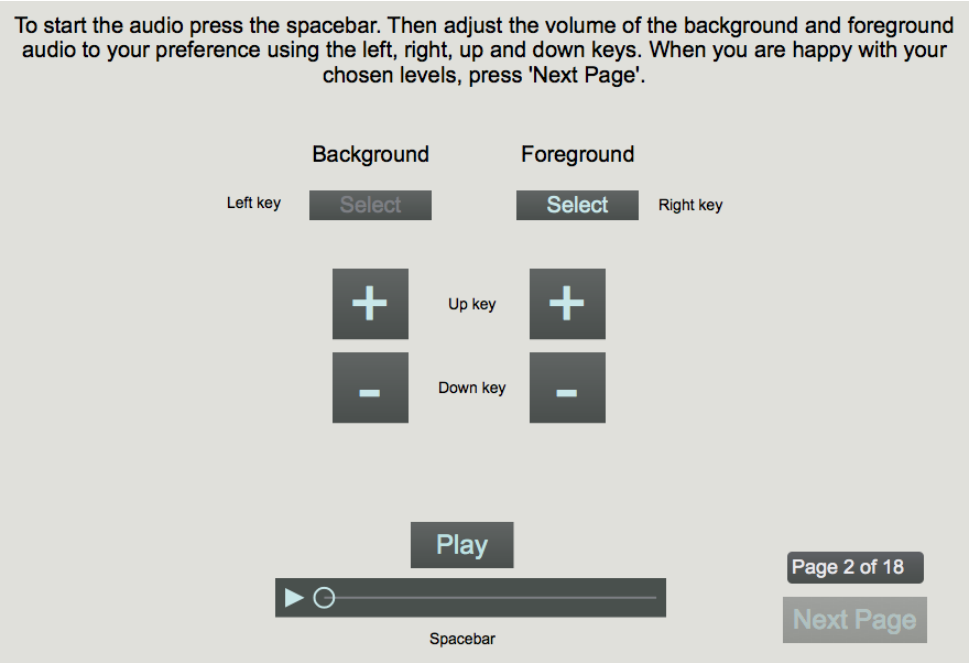


Figure 5.7: Graphical user interface.

white noise, the headphones in this study had approximately 8 dB more attenuation than those in study I (see response "AKG open" in Figure 5.2). However, when comparing attenuation differences using the specific environmental noise recordings, the attenuation differences were only 0.7 dB and 1.9 dB (unweighted, time-integrated) for the "Train" noise and "Crowd" noise respectively.

The remainder of the setup (room, loudspeaker array and curtain) were as presented in Section 5.2.4.

5.4.4 *Procedure*

The procedure was similar to that outlined in Section 5.2.5. From comments made in the first study, the interface was modified by replacing ratio and level controls with a control for background level and a control for foreground level, see Figure 5.7. Again, 36 adjustments were made per participant split into two sessions of 18 with a short break in between. Three environmental noise conditions ("No noise", "Café 65" and "Café 75") were repeated in order to assess participant consistency.

After the adjustment session, semi-structured interviews were conducted. Topics for discussion included the difficulty of the experiment, how the environmental noise was

perceived, how the environmental noise influenced the mix and the mixing process. More specifically, the core questions were:

- How did you find the experiment in terms of its difficulty?
- How would you describe the environmental noise?
- How do you feel the environmental noise influenced your preferred mix?
- Did you always choose a mix of both foreground and background audio, or for some adjustments did you prefer only foreground or background?
- Do you have any other comments about your experience of the experiment?

These questions were meant as initiators to general discussions on the above topics. These discussions were audio recorded and transcribed for analysis.

5.4.5 *Data Collection*

Despite the modification to the interface, the data collection format was the same as in study I, see Section 5.2.6.

5.4.6 *Participants*

A total of 22 participants (age range: 21-42, mean: 27, gender: 15 male, 7 female) participated in this study, 7 of which had also participated in study I. All participants self-reported normal hearing and were fluent in English. 20 out of the 22 participants could be classed as naïve listeners, that is, they were not professionals in the field of audio and had no or little experience of critical listening tests. As with study I, 77% of participants reported that they listen to audio content from a mobile device with headphones everyday.

5.5 RESULTS: STUDY II

The recorded data consisted of background and foreground levels for each condition. From this, FG-BG ratio could be calculated as well as the overall level (the sum of background and foreground levels).

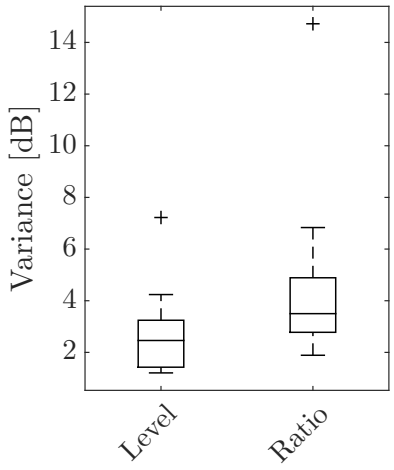
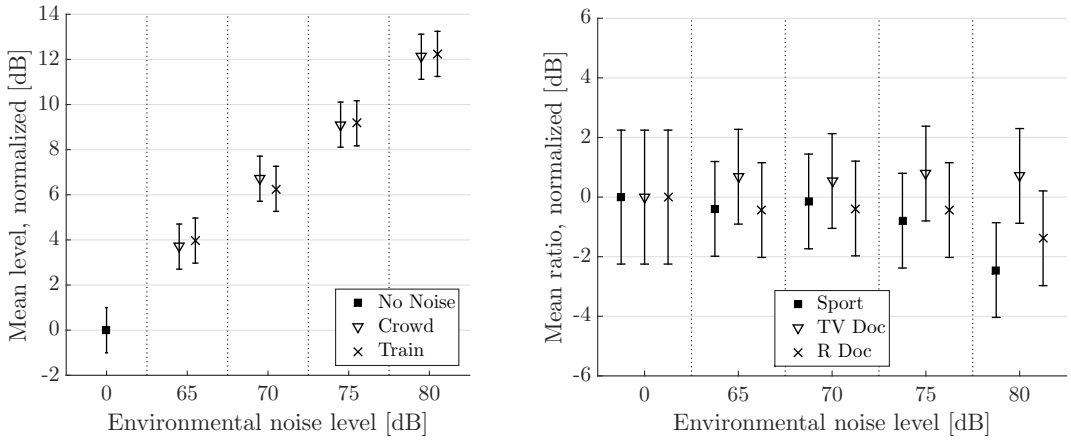


Figure 5.8: Distribution of variance from repeated adjustments for both level and ratio data. The outlying points represent the participants who were excluded from the data analysis.

Participant consistency was checked by calculating the mean variance between repeated adjustments. For each participant, a mean variance was calculated for both the ratio and overall level data. From this analysis it was seen that the mean level variance was 2.6 dB and the mean ratio variance was 4.2 dB. Two participants had outlying variance data, as shown in Figure 5.8, and these participants were therefore excluded from further analysis. Furthermore, a third participant was excluded for adjusting the background or foreground audio to the maximum possible levels on several occasions, which could therefore compromise the preferred ratio data.

5.5.1 *Level*

For the analysis of the level data, a linear mixed model was used. As fixed effects in the model, variables Content, Noise type and Noise level including all interactions were used. To account for differences between individuals, variable Participant was used as a random effect in the model, including intercepts. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. Type III tests of fixed effects revealed that the main effects of Content [$F(2,486) = 15.029$, $p < .001$] and Noise level [$F(3,486) = 98.849$, $p < .001$] were statistically significant ($p < .05$). The remaining effects and interactions were not significant. Therefore the level of the environmental noise significantly influenced the listening level as did the



(a) Level (b) Ratio

Figure 5.9: Mean level (a) and ratio (b) with respect to environmental noise, normalized to the mean of the "No noise" item. A positive ratio represents increased foreground levels whereas a negative ratio represents increased background levels. Error bars show 95% confidence intervals.

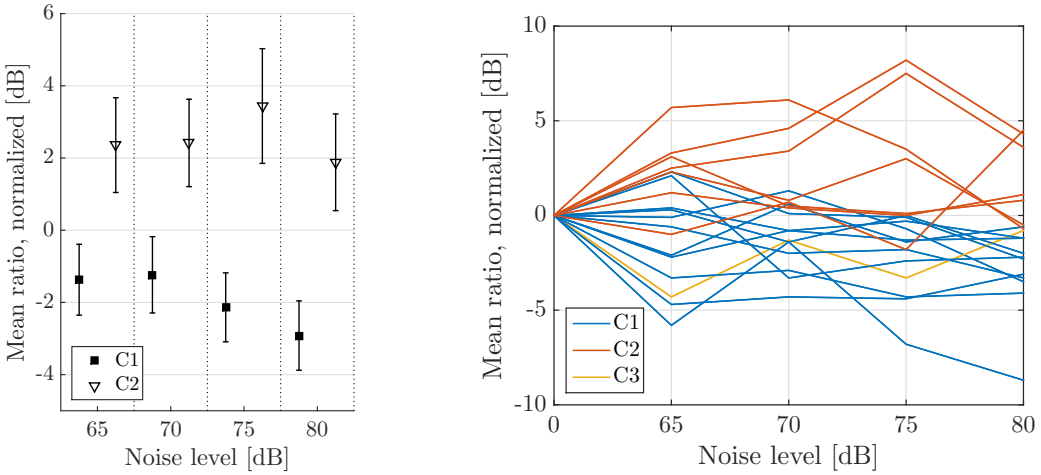
audio item, as shown in Figure 5.9. The type of environmental noise however, did not influence the preferred listening level. This is consistent with study I.

5.5.2 *Ratio*

For FG-BG ratio as the dependent variable in the linear mixed model, Content was the only significant effect [$F(2, 486) = 42.164$, $p < .001$]. When investigating this further it was seen that all three pieces of content have mean ratios significantly different from one another. Most noticeably, content "TV Doc" was mixed with the foreground 4-5 dB louder than the other two items. Unlike study I, environmental noise did not have a significant influence on FG-BG ratio. This can be seen in Figure 5.9.

5.5.3 *Cluster Analysis*

To further explore the variance in the ratio data, a k-means clustering algorithm was applied to the normalized ratio data as a whole with a simple Euclidean distance measure, as in study I. When using two clusters, the data was split into one cluster containing 18 participants and another containing one participant. With three clusters, this changed to clusters containing 11, 7 and one participant. The ratio data from



(a) Mean ratio for clusters C1 and C2. Error bars show 95% confidence intervals.

(b) Mean ratio for each participant. Line colour represents cluster membership.

Figure 5.10: Mean ratio (normalized to the "No noise" item) with respect to environmental noise and cluster membership, averaged by cluster (a) and for each participant individually (b).

these clusters are presented in Figure 5.10. "C1" , "C2" and "C3" represent clusters with 11, 7 and one participants respectively. In C1 a trend is seen towards increased background levels at higher environmental noise levels, whereas the FG-BG ratios in C2 are towards higher foreground levels. With respect to mean ratio versus noise level, C3 appears to be similar to C1, i.e. increased background levels. The different characteristics between C1 and C2, along with the proportions of participants found in each cluster, are consistent with the results from the first study.

5.5.4 *Nature of Ratio Adjustments*

FG-BG ratio adjustments can result from increasing or decreasing either foreground or background levels. To gain more insight into the nature of the ratio adjustments, background level and foreground level were plotted with respect to FG-BG ratio. A linear regression was then applied to these plots, the results of which are presented in Figure 5.11. It is seen that foreground levels are more constant with respect to ratio than background levels. At high environmental noise levels this is even more prominent. This suggests that ratio adjustments are primarily a result of changing background

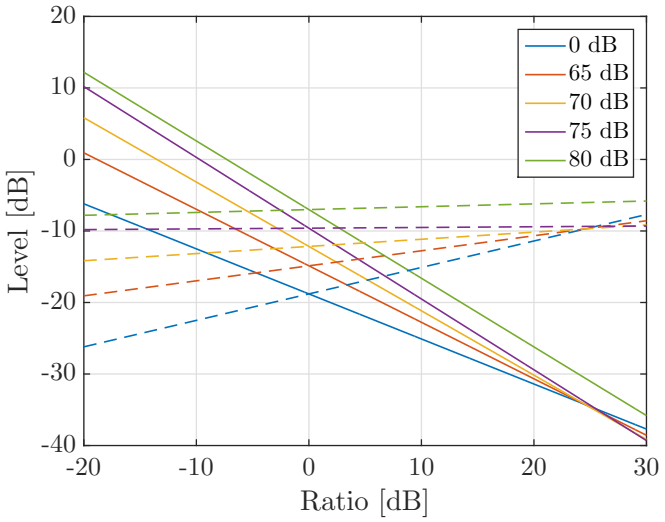


Figure 5.11: Component level vs. ratio for different noise levels. Dashed lines represent FG levels, solid lines represent BG levels.

levels, especially at high environmental noise levels. The reason foreground levels are relatively constant with respect to ratio is likely due to dialogue intelligibility issues.

5.5.5 *Semi-Structured Interviews*

The qualitative data from the semi-structured interviews are considered with qualitative data from study III in Section 5.8.

5.5.6 *Discussion*

The aim of the second study was to further investigate the relationship between environmental noise level and preferred BG-FG audio balance. Unlike study I, FG-BG ratio was not statistically significant with respect to the environmental noise conditions. It is believed that this is due to the large variance in the ratio data, caused by the range of preferences highlighted by the clustering. The clustered ratio data was similar to that from study I. From an analysis of the nature of the ratio adjustments, it was seen that at high environmental noise levels ratio adjustments are predominantly due to changes in the levels of the background components. Semi-structured interviews were conducted and results from this are discussed in Section 5.8.

5.6 EXPERIMENTAL DESIGN: STUDY III

Whereas the results from study I indicated that environmental noise does have an influence on preferred BG-FG balance, those from study II are somewhat inconclusive due to large variances. This is possibly due to the method of adjustment and the relatively small sample size. In study III, a web-based listening test with a multiple comparison method was conducted with the aim of reducing this variance. As the test was web-based a larger sample size could be achieved and the method used meant that the task was simpler for participants. A web-based approach was deemed suitable for this experiment as the differences between conditions were perceptually quite large, and therefore strict laboratory reproduction conditions were not necessary. On the other hand, limitations of a web-based approach include a lack of control over the participants' environmental conditions as well as the possible variety in reproduction setups used (sound cards, headphone models etc.). By asking participants to conduct the experiment in a quiet environment and to be consistent with their reproduction setup, these limitations can however be minimised sufficiently.

On each page of the web-based interface, participants listened to five mixes of the same audio content (with different FG-BG ratios) and were required to select their most preferred mix. This was done with and without environmental noise mixed into the audio files. The study consisted of two sessions - the second being optional. In each of these sessions one piece of audio content and one type of environmental noise were used. This design minimised the duration of the study so that each session took approximately 10 minutes to complete. Specifics of these variables are outlined in the following sections.

5.6.1 *Audio Excerpts*

Two audio excerpts were used in this study: "Sport" and "Doc". These were of a similar nature to the previous sport and documentary clips described. The duration of these were 24 s and 16 s respectively. Mixes of these two items were made with FG-BG ratios of ± 9, ± 4.5 and 0 dB. These ratios were decided upon from a combination of examining the previous results from studies I and II and choosing ratios that the majority of untrained listeners should be able to differentiate between. It should be noted that these ratios were achieved by keeping the level of the foreground components constant and adjusting the level of the background components. The reasoning

behind this was that at high environmental noise levels in the previous study, the change in FG-BG ratios came predominantly from changing background component levels, as discussed in Section 5.5.4.

5.6.1.1 *Environmental Noise*

The two environmental noise clips used in study I were also used in this study. The "Café" clip was used in combination with the "Sport" content and the "Train" clip was used in combination with the "Doc" content. The environmental noise was trimmed to match the duration of the audio excerpts. As participants' listening levels could not be calibrated, instead of a fixed absolute level the environmental noise level was set as a fixed signal-to-noise ratio (SNR) in relation to the audio excerpts. The appropriate SNR values were calculated from the mean listening levels for the 80 dB environmental noise condition in study II, taking into account headphone attenuation of the environmental noise. The SNR used was 2.3 dB.

Whereas the previous two studies delivered the environmental noise through loudspeakers, the web-based nature of this study meant that the environmental noise was delivered through headphones. In order to reproduce the spatial information from the environmental noise, the ambisonic reproduction of the noise clips was recorded binaurally using a Neumann KU 100 dummy head. Additionally, AKG K702 open back headphones were placed on the dummy head in order to include the attenuation effects of the headphones, which were present in the previous studies. The binaural environmental noise was embedded into the audio content files at the relevant SNR.

5.6.2 *Procedure*

The study was implemented using the Web Audio Evaluation Tool (Jillings et al., 2015), a browser-based listening test environment based on the HTML5 Web Audio API. After reading the introduction and instructions, participants were presented with a familiarisation page in which a sample of the foreground content (i.e. dialogue) was played. Participants were asked to adjust their listening level to a comfortable volume and it was additionally stated that, if possible, they should keep this level constant throughout the test. The subsequent four pages were the rating pages; without environmental noise and with environmental noise pages plus repeats. Instructions to the participants were "Imagine you are at home watching a nature documentary. Switch between the mixes below and select the mix that you would most prefer in this situa-

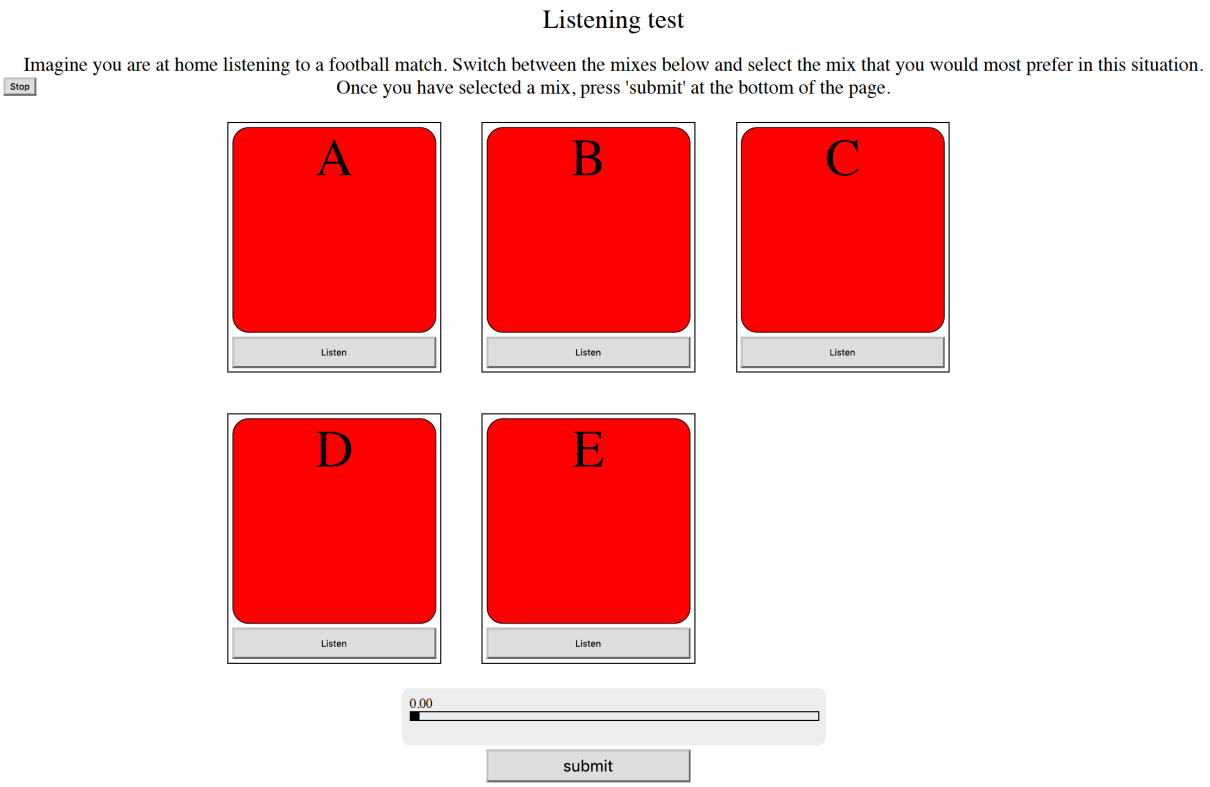
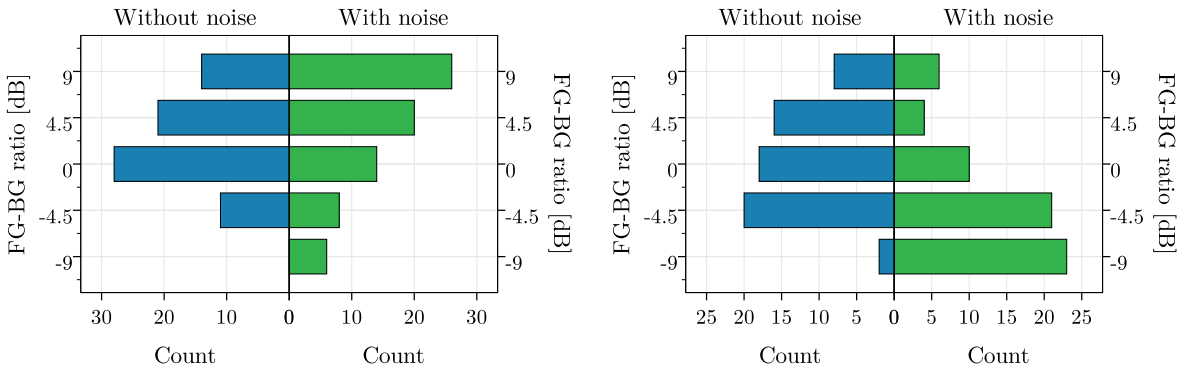


Figure 5.12: Graphical user interface.

tion" with the content and environment descriptions changed accordingly. For example, for the "Sport" content in combination with the "Café" noise, the instructions read "Imagine you are in a café listening to a football match on your phone...". After the rating pages participants were asked several qualitative questions including "Do you feel that the café/train noise influenced your preferred mix?", "If so, in what way?" and "Please enter any comments about why you think you changed your preferred mix".

5.6.3 *Data Collection*

For each participant, an Extensible Markup Language (XML) results file was dynamically generated by the interface upon clicking the 'Submit' button. The data contained in this results file included all responses (both mix selections and qualitative responses) as well as usage data, such as participant's browser and operating system, timestamp, test timer, element timer and element tracker. This usage data was not considered in the analysis.



(a) Sport (b) Doc

Figure 5.13: Histograms of chosen mix counts for without noise and with noise conditions, for the content (a) Sport and (b) Doc.

5.6.4 *Participants*

50 participants completed the first session of the study with 37 of these going on to complete the optional second session. Demographic data was not collected so as to minimise personal data collection and also to reduce the duration of the study. Participants were recruited from a range of platforms including social media and company mailing lists.

5.7 RESULTS: STUDY III

Participant variance was analysed by comparing repeat choices for each condition (without/with noise). It was decided to exclude participants with a repeat variation of greater than 4.5 dB for either condition (4.5 dB being the smallest step in FG-BG ratio). For "Sport", 13 participants (26%) were excluded with 37 remaining, and for "Doc", five participants (14%) were excluded with 32 remaining.

To examine the effect of environmental noise on preferred FG-BG ratio, choice histograms are plotted which show the times each mix was chosen for each condition, Figure 5.13. For "Sport", without environmental noise the most chosen mix was 0 dB (FG-BG ratio) with a normal distribution around this. In the with noise condition however, the most chosen mix was 9 dB, i.e. the mix with the lowest level of background components. It is also seen that the mix with the highest level of background components, −9 dB, was not chosen in the without noise condition, although in the with noise condition is was. Therefore the majority of participants adjusted their mix

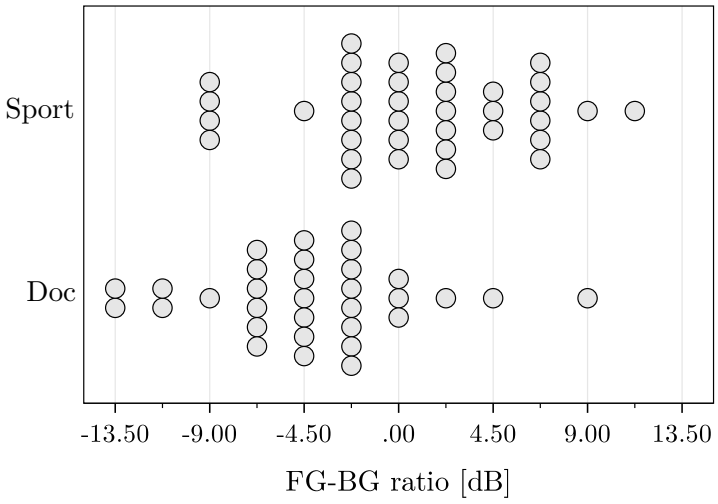


Figure 5.14: Dot plots of the mean difference between without noise choices to with noise choices for each participant.

to lower background levels in the with noise condition but some participants increase the mix to higher background levels.

For the documentary content, without environmental noise the most chosen mix was −4.5 dB. Despite the preference being for high background audio levels, very few participants chose the −9 dB mix, with every other mix being more popular. With environmental noise however, the −9 dB mix was the most preferred. This is opposite to the behaviour of the sport content which shows that the content plays a big part in how environmental noise influences the preferred mix.

To quantify the significance of the differences in the without noise/with noise choice distributions, a McNemar-Bowker Chi-square test was used. In this case, the null hypothesis is that the distribution of choices is equal in the without noise condition and the with noise condition. The alternative hypothesis is therefore that there is a difference in distribution of choices for the two conditions. Both the "Sport" $[\chi^2(9, N = 74) = 18.82, p = 0.027]$ and the "Doc" $[\chi^2(9, N = 64) = 38.11, p < 0.001]$ content show significantly different without noise and with noise choice distributions.

To consider the results on an individual basis, i.e. the difference between conditions for each participant, dot plots are presented which show the mean difference between without noise choices to with noise choices for each participant, Figure 5.14. As a variation of 4.5 dB was allowed between repeats and the points in Figure 5.14 represent the mean between two choices, each point has a possible variance of 2.25 dB. Therefore, points within ±2.25 dB of 0 dB have error margins spanning the three outcomes of increased background, no change and increased foreground. Despite this, the over-

all trends in this figure highlight what was seen in the previous choice histograms. For "Sport", the trend is towards lower background audio levels, although with a group of participants who preferred an increase in background levels. For "Doc", the trend is towards higher background audio levels, but this is not the case for all participants.

The responses from the post-test surveys are analysed in the following section, along with the qualitative data from study II.

By conducting this study as a web-based experiment, it was possible to gather both quantitative and qualitative responses from a relatively large number of participants. This, along with the adjusted method, meant that trends seen in the ratio data were more clear than in the previous studies, reinforcing and extending the previous results. As with the previous results, results from this study have shown that environmental noise can influence preferred background-foreground audio balance. The choice histograms presented highlight that the preferred mix in the presence of noise is very much dependent upon the audio content. As in the previous studies, it was seen that participant responses should not be generalised.

5.8 QUALITATIVE ANALYSIS

In order to gain insight into why participants changed their preferred mixes in the presence of environmental noise, the qualitative data from studies II and III are now considered. In particular, we consider responses to the question "How do you feel the environmental noise influenced your preferred mix?" from study II and "Please enter any comments about why you think you changed your preferred mix" for both content items in study III. A total of 88 responses are considered. The interviews from study II were audio recorded so the first step in the analysis was to transcribe these recordings. This was done using the software NVivo. In the case of study III, the responses were typed so no transcription was necessary. Analysis of the data was based on a thematic approach (Braun and Clarke, 2006), i.e. organising sections of the data into recurrent themes.

Figure 5.15 shows a schematic diagram representing the main themes identified that relate to adjusting the mix due to environmental noise. As can be seen, these themes are split into those related to increasing background components relative to foreground components and those related to increasing foreground components relative to background components.

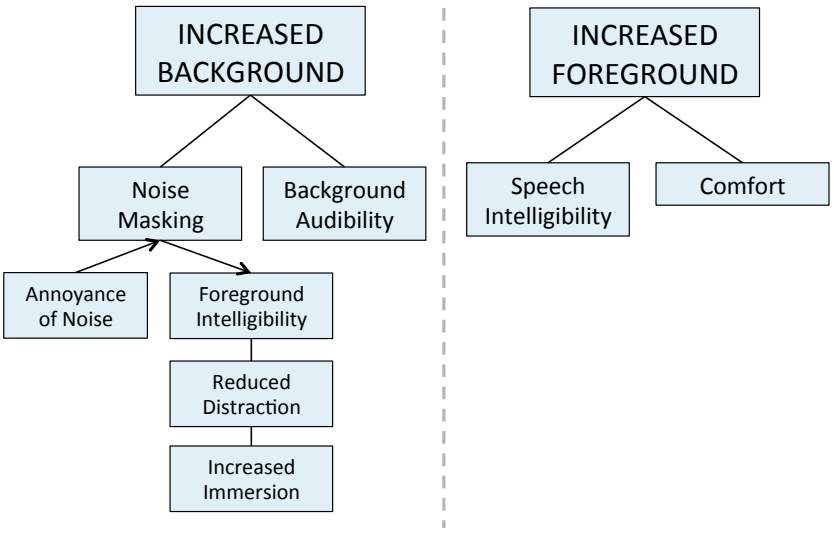


Figure 5.15: Main themes identified from thematic analysis related to adjusting an audio mix due to environmental noise.

5.8.1 *Increased Background*

35 responses were related to increasing the background components of the mix relative to the foreground. A large proportion of these were comments related to increasing the background in order to mask the environmental noise.

> "I think sometimes I consciously went higher with the background noise to drown out the environmental noise."

> "The crowd noise was more acceptable when it helped to mask the café ambient noise."

> "To drown out the background train noise to a greater extent."

Several participants explicitly stated that this was due to the "annoying" nature of the environmental noise. This behaviour of masking the noise is due to the characteristics of the foreground and background components. There are fewer gaps in the background components (e.g. music, crowd) than the foreground components (e.g. speech) and as a result an increased background can help block out the unwanted environmental noise. The results from masking the noise include increased foreground intelligibility:

> "It somehow actually made it easier to listen to the narration by making it more difficult to hear the train",

reduced distraction:

> "The louder and more uniform and relevant background noise of the crowd prevents distraction",

and increased immersion:

> "...it was definitely a case of, I feel I need to bring this up more to drown things out and to feel like I'm in the programme."

Besides from noise masking, another prominent reason for increasing the background was audibility of the background components.

> "I wanted to turn [the] background music up, to be able to hear it."

> "...when it's really noisy it's much harder to get any background so you just like get the pieces that are like kind of high, so I think I put it like higher when it was like really noisy around."

5.8.2 *Increased Foreground*

32 responses were related to increasing the foreground components of the mix relative to the background. Speech intelligibility was the primary reason for this.

> "In a cafe (noisy environment), sound quality and a sense of immersion is secondary to understanding the commentary."

> "I wanted the narration to be clear and so I had to reduce the volume of the music..."

> "...when it was louder I would have the narrative on, the foreground on higher so I could concentrate more on the foreground."

It is apparent that the environmental noise made concentrating on and understanding the foreground speech content more of a challenge.

Another reason for increasing the foreground relative to the background was due to comfort reasons. By keeping the foreground content at an intelligible level and reducing the background components, participants reduced the overall audio level.

> "I know there's information being conveyed in the background sounds and I only wanna lose out on that if it's too noisy to make it comfortable listening to the foreground sounds... in the really noisy environments I wanted the, I guess to keep the additional noise I'm listening to down to a minimum, so that it's comfortable."

> "The café... mixed with the crowd noise of the football match it becomes too noisy."

Additionally, listening comfort increased as there were less competing sounds in the mix.

> "The background cafe noise was really irritating, and having even more noise to add into that was just even worse."

> "The different sounds were off-putting; almost a sensory overload."

5.8.3 *No Change*

A number of responses were related to participants not changing their preferred mix in the presence of environmental noise. In such responses, participants often stated that they liked a mix a certain way (e.g. "...liked the music slightly loud all the time") and that the environmental noise did not influence this. These comments were less revealing and were therefore not coded into further themes.

5.9 DISCUSSION

In the three studies presented above, the influence of environmental noise on preferred BG-FG balance has been explored. In the first study, it was seen that environmental noise can significantly influence FG-BG ratio with the overall trend being towards higher background levels in the presence of environmental noise. Furthermore, the participants were clustered into two groups by their preferences. The largest cluster adjusted the FG-BG ratio to higher background levels in the presence of environmental noise whereas a second cluster preferred unchanged to increased foreground ratios.

The second study explored this further with a larger range of environmental noise stimuli and the addition of semi-structured interviews to gain a qualitative under-

standing of the two clusters identified in study I. Unlike the first study, environmental noise was not seen to have a significant influence on the chosen FG-BG ratio. This lack of significance could possibly be attributed to the individual nature of the ratio adjustments, increasing variation around the mean. This could therefore indicate that the overall ratio trends seen in study I are not generalisable to other populations. The large variance in results was further examined with a cluster analysis and again two main clusters were identified with similar trends as in study I. Additionally, it was seen that at high environmental noise levels, the ratio adjustments were predominantly due to adjustments of the background component levels, that is the foreground levels were approximately constant with respect to ratio.

The third study, which was web-based, aimed to reduce the variance in the data by using a simpler evaluation task and increasing the sample size. As in study I, it was seen that environmental noise significantly influences FG-BG ratio. Moreover, the differences that audio content and environmental noise make on the preferred mix were highlighted. For the sport content the overall trend was to adjust the FG-BG ratio to higher FG levels in the presence of noise, whereas for the documentary content the opposite trend was seen.

The themes identified in the qualitative analysis from studies II and III revealed the different approaches taken by participants to minimise the effect of environmental noise on the overall listening experience. On the one hand, participants chose to increase the background components in order to mask the environmental noise and to ensure that the background components were audible above the noise. This noise masking from the background components in turn increased foreground intelligibility, reduced distraction and increased immersion. On the other hand, participants chose to increase the foreground components in order to improve the speech intelligibility and also the comfort of the overall experience. This dichotomy of qualitative responses is in agreement with the quantitative data.

The nature of the audio content is clearly a big factor in the responses seen. It is the continuous nature of the background components (crowd, music) that enable them to be used to mask the environmental noise. Furthermore, it is the nature of the background components that seems to influence whether the majority of participants increase or decrease their level in the presence of noise. The sport content had background components that were considered as noise and not particularly necessary by some participants. On the other hand, the documentary had background components that were considered less noise-like and more essential to the listening experience.

One limitation of this set of studies is the limited range of content used and therefore the effect of content and content-noise interaction should be further investigated.

Another point of further study could be the relationship between dynamic range adaptation and mix adaptation. Some participants mentioned that they raised the level of the background components in order to clearly hear all of the background components. Such results could also be possible by adjusting the dynamic range of the content to the environmental noise, as in (Mason et al., 2015). The possibility of using both methods in conjunction and their relative contributions to an improved listening experience should be investigated.

The findings from these studies show that with suitable object-based content, a simple background-foreground level control could be beneficial for users listening in noisy environments. As well as improving the listening experience, this could also help to prevent hearing loss; instead of increasing the level of the content as a whole, only the level of the desired components could be increased, therefore reducing exposure. Context-aware adaptation is ultimately desirable, although the influence of content and participant on the preferred mix means that this is not a trivial task.

In terms of the study of context IFs in general, the results presented here highlight the need to consider context IFs when evaluating certain aspects of next generation audio. The quantitative and qualitative data collected suggests that by considering context IFs, an improved listening experience can be delivered to the user. By considering other context IFs, an understanding can be formed of how best to satisfy the needs of users in a range of contexts. Developers of technology and providers of content could therefore utilise this information to provide services with a higher QoE.

Whereas this study only considered the context IF of environmental noise, there are a range of other factors that could influence the listening experience of next generation audio, for example those outlined in Table 3.1. For some of these context IFs, aspects of the methodology used for this particular case study could be suitable for their investigation. A laboratory-based approach, whereby context IFs are reproduced in an isolated manner, proved useful for initial investigations of possible effects. Certain temporal, economic and task based context IFs (e.g. frequency of use, brand and multitasking) could possibly also be investigated in laboratory-based studies. For other physical and social context IFs (e.g. location and interpersonal actions) however, such approaches may be less suitable. Instead, in-the-wild approaches might be more suitable, in which case it is important to understand and report the various factors that make up a given context, for example by using the method proposed by Jumisko-

Pyykkö and Utriainen, (2011). Environmental noise was an ideal context IF to study initially due to the flexibility in possible research approaches (i.e. laboratory-based, web-based, in-the-wild). A complementary follow-up study might be one to investigate a context IF that is not suited to laboratory-based approaches, so as to highlight possible methodologies for other such context IFs.

In terms of studying content adaptation via object-based audio, the method of adjustment combined with semi-structured interviews proved useful for exploratory investigations and might therefore be suitable for investigating the adaptation of content in other manners, such as adapting its complexity and spatial characteristics. However, it also produced results with a high variance which led to the use of a multiple stimulus approach. Combing adjustment tasks with multiple stimulus or satisfaction tasks, such as in the "adjustment / satisfaction" method recently proposed by Torcoli et al., (2017), could be useful for further investigations on content adaptation.

5.10 SUMMARY

In this chapter, we have seen how next generation audio can be utilised to improve QoE in relation to context influence factors. This was achieved by means of three studies investigating if environmental noise influences preferred audio object balance. Both quantitative and qualitative methods revealed that environmental noise does indeed influence preferred audio object balance and the reasons behind this were explored through thematic analysis. The results presented highlight the potential of considering context IFs when designing for next generation audio experiences. Strengths and weaknesses of the methods used in these studies were discussed and it was seen that certain aspects of the methods may be beneficial for further studies on context IFs.

Part III

HUMAN

6

THE ROLE OF HUMAN FACTORS ON OVERALL LISTENING EXPERIENCE

6.1 INTRODUCTION

For the third part of this thesis, the role of human influence factors on the QoE of next generation audio are investigated. As previously discussed, human IFs are properties or characteristics of a human user. Human IFs can be low-level, such as those related to the physical, emotional and mental constitution of the user, or high-level, such as socio-economic situation, education background and values. As discussed in Section 3.3.3, human influence factors are not particularly well studied in the field of audio quality evaluation. Typically, studies only distinguish between listeners on the basis of experience (i.e. naïve, inexperienced, experienced and so on), with few studies assessing other user related variables. However, with next generation audio offering personalised experiences to the user, it is increasingly important to consider how variables related to the individual can influence the perceived QoE.

One evaluation measure that has been used to illustrate differences in listener type is "overall listening experience" (OLE), as introduced in Section 2.5.4. This is an affective measure that is intended to include all possible factors that may influence listeners' ratings of stimuli, for example the song, lyrics, technical audio quality, the listener's mood and the reproduction system. As with QoE, by it's nature OLE is therefore user (or listener) dependent. This was highlighted in a study which showed that the relative influence of content and technical quality on OLE depends on the individual; on the one hand some users are heavily influenced by content when making OLE judgements and, on the other hand, some users are heavily influenced by technical audio quality when making OLE judgements, with a continuum of users between (Schoeffler and Herre, 2014).

In order to tailor systems and services to the appropriate audience, it would be beneficial to know what types of listeners are using the services and systems in question. For example, if it was known that the overall listening experience of a certain user group is highly influenced by technical audio quality, it would be desirable to pro-

vide them with the best quality available. Likewise, if it was known that the overall listening experience of a different user group is highly influenced by the content, it would be less problematic if the quality was reduced.

In this chapter, an experimental study is presented with the aim of identifying psychographic[1] variables that significantly influence whether a listener is heavily influenced by content or quality when making OLE ratings. As well as having direct applications, such as those mentioned above, this study provides insight into how human factors can influence quality of experience in general.

In order to investigate the above aim, it is necessary to present audio items of various quality levels to the participants. As such, this experiment was used as an opportunity to investigate a secondary objective. The influence of various systems and technologies on OLE have previously been investigated, for instance single/multichannel systems and 3D audio systems. However, the influence of binaural processing on OLE had yet to be studied. As was previously discussed, binaural audio is perhaps one of the most accessible forms of immersive audio as the majority of consumers already own the technology for its playback. This, coupled with the popularity of headphone listening, means that binaural content is becoming increasingly available to the audience. The majority of studies on the evaluation of binaural reproduction focus on specific technical aspects and are typically conducted in laboratory settings. Such studies have a high internal validity and are indeed highly relevant for the advancement of the technology. However, it is also important to evaluate the benefits of binaural audio with a more holistic, QoE-based approach. Thus, the evaluation of the influence of binaural processing on OLE is a worthwhile aim, and it is this which is the secondary objective of the study presented in this chapter.

6.2 EXPERIMENTAL DESIGN

This experiment was conducted as a web-based study. This was seen as appropriate as a large number of participants were required from a range of backgrounds and, furthermore, the differences between stimuli were not so small as to necessitate strict laboratory reproduction conditions. Moreover, a web-based approach leads to a higher external validity, which is an important consideration when evaluating quality of experience; participants listened to the content in a situation typical of their normal

1 *Pyschographics* can be defined as "The study and classification of people according to their attitudes, aspirations, and other psychological criteria..." (OED Online, 2017).

listening environment and with the technology that they would typically use. As previously discussed, web-based studies are limited by the associated lack of control of certain variables, such as the participants' environmental conditions and reproduction setups, and it is therefore important to minimise these risks through instructions.

The study was split into three sections; an online questionnaire to collect psychographic data and two online listening sessions. These are described in more detail in the following sections.

6.2.1 *Psychographic Data Collection*

The psychographic data were collected by means of an online questionnaire, the overall form of which was inspired by (Jumisko-Pyykkö and Häkkinen, 2008). The data collected can be roughly categorised into groups relating to demographics, experience and attitudes towards audio technology. Additionally, name and email were collected during each session for identification purposes.

Demographics

Data collected relating to demographics includes gender, age group, level of education (British system) and self-reported hearing normality. More specifically, age group and level of education were categorised as:

- Age group
  - 17 or younger
  - 18 - 25
  - 26 - 35
  - 36 - 45
  - 46 - 55
  - 56 - 65
  - 66 or older
- Level of education
  - Compulsory education
  - College, sixth form or equivalent
  - Undergraduate degree
  - Postgraduate degree
  - None of the above

Experience

To assess experience in the field of audio technology and specific experience relating to headphone usage and binaural audio experience, the following four questions were used.

- Select the statement that best describes the role of audio technology in your work and hobbies:
  - I study or work mainly in the field of audio technology
  - My work or hobbies involve some knowledge of audio technology
  - My work or hobbies are not related to audio technology
- Select the statement that best describes your headphone listening habits:
  - I listen to audio over headphones most days
  - I often listen to audio over headphones
  - I rarely listen to audio over headphones
  - I never listen to audio over headphones
- Select the statement that best describes your experience with binaural audio:
  - I have no experience of listening to binaural audio
  - I have limited experience of listening to binaural audio
  - I am experienced in listening to binaural audio
  - I'm not sure
- How many listening experiments have you previously participated in?
  - None
  - 1-5
  - 6-10
  - More than 10

Attitudes Towards Audio Technology

A combination of two previously reported questionnaires was used to measure attitudes towards audio technology. The first of these, The Domain Specific Innovativeness (DSI) scale (Goldsmith and Hofacker, 1991), has previously been used in a range

of fields to measure consumer innovativeness[1], including studies related to quality assessment of mobile television (Jumisko-Pyykkö and Häkkinen, 2008). In addition to this scale, parts of a questionnaire designed to measure technical affinity, known as the TA-EG (Karrer et al., 2009), were used to measure competence and enthusiasm. This questionnaire was originally designed for use with German speakers and was therefore translated for this study. In the original TA-EG questionnaire there are also measures of "negative attitudes" and "positive attitudes" towards technology in general. These items did not translate well to the specific case of audio technology and were therefore not included. The complete list of statements to measure attitudes towards audio technology is as follows:

- Competence
  - I know most functions on the audio devices I own
  - *I struggle/would struggle to understand audio technology magazines*
  - I find it easy learning how to operate audio devices
  - I'm well versed in the field of audio technology
- Enthusiasm
  - I stay informed about audio technology, even if I don't intend to make a purchase
  - I love owning new audio technology
  - I get excited when a new device related to audio technology is brought to market
  - I like to go into specialist retailers for audio technology
  - I enjoy trying out audio technology
- Domain Specific Innovativness
  - *In general, I am among the last in my circle of friends to buy new audio technology when it appears*
  - If I heard that a new item of audio technology was available to purchase, I would be interested enough to buy it
  - *Compared to my friends I don't own much audio technology*

1 Domain specific innovativeness reflects the tendency to learn about and adopt innovations (new products) within a specific domain of interest (Goldsmith and Hofacker, 1991), based on (Midgley and Dowling, 1978).

- *In general, I am the last in my circle of friends to know about the latest audio technology*
- I will not buy new audio technology if I haven't tried it yet
- I like to buy new audio technology before other people do

These claims were presented in a continuous list without the headings shown above. Participants were instructed to "rate your attitude towards audio technology with the following statements" with ratings being made on a five-point Likert scale ranging from "strongly disagree" to "strongly agree". The statements in italic type are negatively phrased and the corresponding scores must therefore be reversed for analysis.

6.2.2 *Stimuli*

10 music items were used for the main rating sessions with an additional four being used for the familiarisation pages, see Table 6.1. These spanned a range of genres and suitable phrases were selected that ranged in duration from 16 - 25 seconds (mean 21.9 s). The main selection criterion for these items was that they were available in formats that were suitable for the generation of binaural versions, i.e. captured with appropriate microphone techniques for the live classical and jazz performances and available as multitrack recordings for the popular items. Further criteria were that they were relatively broadband in nature, had a relatively wide stereo image and would elicit a range of preferences.

For each item four conditions were created: stereo, mono, 3.5 kHz low-pass filtered and binaural. All items were available as stereo mixes and these were used as the basis for the creation of the spatially degraded mono and timbrally degraded 3.5 kHz low-pass conditions.[1] The mono items were created by passively downmixing the stereo items in accordance with ITU-R BS.775 (2012a). The 3.5 kHz low-passed items were generated with a 5th-order Butterworth filter. A professional sound engineer experienced in mixing spatial audio assisted in the generation of the binaural items. The classical and jazz items were captured with a Schoeps ORTF-3D microphone array plus a range of close mics and were binaurally post-processed using a BBC R&D binaural renderer. For an informal description of the recording process please

1 It should be noted that previous studies have reported low-pass filtering to also cause small deteriorations in spatial quality and down-mixing to cause small deteriorations in timbral quality (Zielinski et al., 2005). As this effect is small, for the purposes of this study it was assumed that down-mixing only caused spatial degradation and low-pass filtering only caused timbral degradation.

Table 6.1: Overview of the content items used. Starred items were used in the familiarisation stage only.

Genre	Artist	Title	Duration	Notes
Classical - Choral	Bach	Komm, Jesu, Komm	21 s	Performed by The Sixteen for BBC Prom 42, 2016
Jazz - Big Band	Duke Ellington	Circle of Fourths	23 s	Performed by the National Youth Jazz Orchestra of Scotland for BBC Prom 28, 2016
Jazz - Trumpet Improv.	Duke Ellington	Lady Mac	24 s	Performed by the National Youth Jazz Orchestra of Scotland for BBC Prom 28, 2016
Folk	Hezekiah Jones	Borrowed Heart	25 s	Recorded for Weathervane Music's Shaking Through, Vol. 2, Ep. 4
Indie	Hop Along	Sister Cities	22 s	Recorded for Weathervane Music's Shaking Through, Vol. 4, Ep. 5
Electronic	La Big Vic	Musica	18 s	Recorded for Weathervane Music's Shaking Through Vol. 2, Ep. 3
Hip Hop	Lushlife	Toynbee Suite	25 s	Recorded for Weathervane Music's Shaking Through Vol. 4, Ep. 8
Classical - Orchestral	Prokofiev	Romeo and Juliet	23 s	Performed by the BBC National Orchestra of Wales for BBC Prom 16, 2016
Classical - Orchestral	Schubert	Symphony No. 9	23 s	Performed by the BBC Philharmonic for BBC Prom 24, 2016
Pop	Steven A. Clarke	Bounty	20 s	Recorded for Weathervane Music's Shaking Through Vol. 4, Ep. 1
Folk*	Lea Thomas	Wild As You Are	20 s	Recorded for Weathervane Music's Shaking Through Vol. 8, Ep. 1
Classical - Orchestral*	Schubert	Symphony No. 9	24 s	Performed by the BBC Philharmonic for BBC Prom 24, 2016
Classical - Orchestral*	Tchaikovsky	Romeo and Juliet	16 s	Performed by the BBC Symphony Orchestra for BBC Prom 1, 2016
Indie*	The Tontons	Lush	23 s	Recorded for Weathervane Music's Shaking Through Vol. 5, Ep. 2

refer to (BBC R&D, Tom Parnell, 2017). It should be noted that the same engineer produced both the stereo and binaural mixes for the classical and jazz items. With the remaining popular items, multitrack recordings were available which included a combination of individual instrument tracks and grouped instrument tracks. These tracks were treated as audio objects and were binaurally post-processed using the same software as the classical and jazz items.

All stimuli had a 250 ms fade-in and fade-out applied and were presented as 44.1 kHz / 16 bit WAV files. Additionally, a two-stage loudness alignment process was conducted to equalise the loudness of all stimuli. The first stage involved aligning all stereo items to a target loudness of −18 LUFS in accordance with (ITU-R, 2015d). A target loudness of −18 LUFS was chosen as such a level is more appropriate for mobile devices than the more typical −23 LUFS (AES, 2015). Secondly, the remaining conditions for each item were aligned to the loudness of the stereo condition using the Glasberg and Moore loudness model applicable to time-varying sounds (Glasberg and Moore, 2002). For each item, loudness values for each stereo channel were calculated individually and then these two values were averaged to produce a single loudness value. Furthermore, the model was applied without an outer ear transfer function stage as the stimuli were to be presented over headphones.

6.2.3 *Procedure of Listening Sessions*

The listening sessions were conducted online by means of the software webMUSHRA (Schoeffler et al., 2015b). Each participant completed two listening sessions with a duration of approximately 15-20 minutes each. These were separated by a break of at least one week so as to prevent over familiarisation of the stimuli which could lead to annoyance and bias in the ratings. Each listening session included an introduction page, a familiarisation page, a multiple stimuli page and 20 single stimulus pages. Both of the two sessions were identical apart from the stimuli used in the single stimulus ratings.

On the introduction page participants were welcomed and asked to ensure that they were in a quiet space with headphones plugged into their device. The following instructions were given about the task to be completed:

> "In this experiment you will listen to various excerpts of music. For each excerpt you will be asked to rate your overall listening experience on a simple scale. In particular, you will be asked 'How much do you enjoy

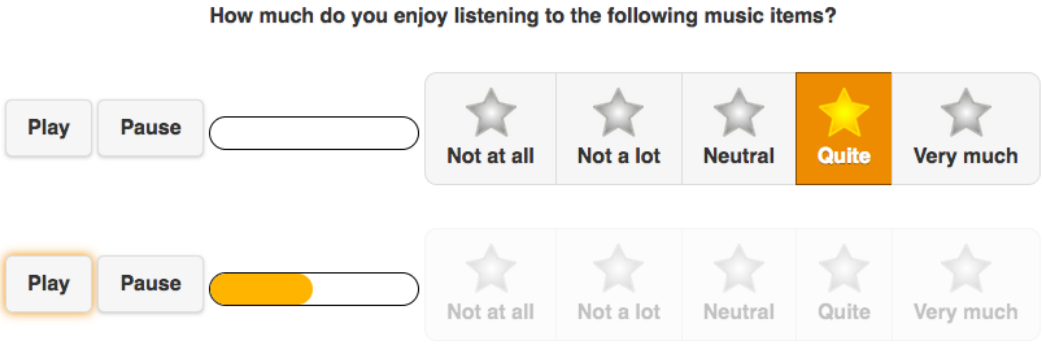


Figure 6.1: User interface for OLE ratings.

listening to the following music item(s)' with possible answers ranging from 'not at all' to 'very much'. When making your ratings you should take everything into account that you would normally in a real-world scenario (e.g. your taste in music, the audio quality etc.)."

It should be noted that here the term "overall listening experience" is expressed as a rating of "enjoyment", as in previous implementations of the OLE method. This follows the assumption that for music consumption the overall experience can be represented by "enjoyment" alone. Considering the definition of QoE (Möller and Raake, 2014) (based on (Le Callet, Möller, and Perkis, 2013)), which refers to the fulfilment of expectations and needs with respect to "utility and/or enjoyment", this assumption is sensible as utility in this context is not relevant. For further discussions on the choice of question when assessing OLE please refer to Schoeffler, (2017) (pp. 35-42).

After the introduction, a familiarisation page allowed participants to play and rate four stimuli in order to adjust the volume of their device to a comfortable level and to practice using the interface. It was stated that once adjusted, the volume should not be changed during the remainder of the experiment. The four stimuli included one of each quality condition and were not used in the main rating pages. As with all of the rating pages, the order of the stimuli on the page was randomised. Ratings were made on a five-star Likert scale with labels of "not at all", "not a lot", "neutral", "quite" and "very much", see Figure 6.1. Before making a rating of an item, participants had to listen to the item completely and before moving on to the next page, all items had to be rated. After the familiarisation page, participants made ratings of all of the stereo stimuli (10 items in total) presented on a single page, so as to reduce floor and ceiling effects (Schoeffler, 2017). These ratings are known as the basic item ratings (BIRs).

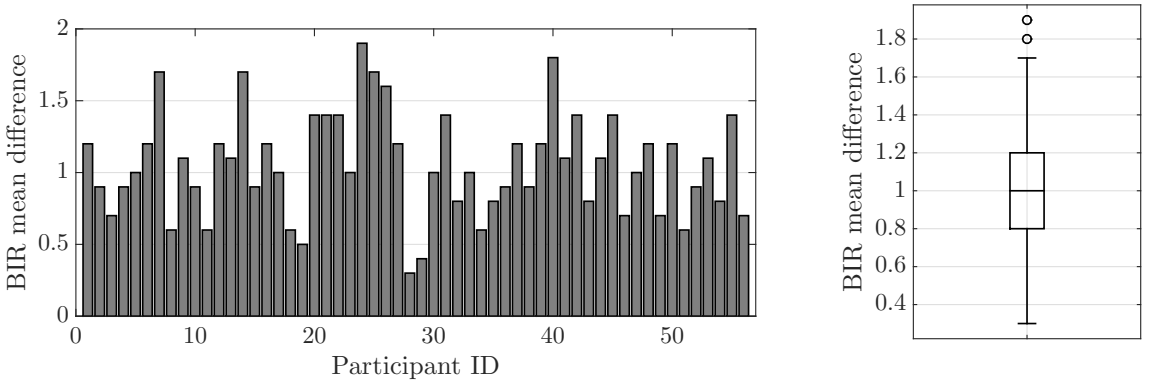
Following this, single stimulus ratings were made for 20 stimuli which consisted of each content item at two quality levels. These are given through a single-stimulus procedure as such an approach is more representative of real-world listening scenarios (Schoeffler, 2017). Each quality level appeared the same number of times in each session. Over the two sessions, participants therefore rated all stimuli (10 items by four conditions) by the single stimulus method. These are known as item ratings (IRs). The allocation of stimuli to sessions was predetermined. To ensure that all combinations of quality levels for each content item were included, six configurations were needed and the assignment of these to participants was balanced.

6.2.4 *Data Collection*

Upon completion of a session, data was appended to a CSV file. Fields recorded were session ID (to identify stimuli configuration), participant name, participant email, trial ID (e.g. familiarisation / BIR / IR), stimulus rating, stimulus name, rating time and submission timestamp.

6.2.5 *Participants*

Participants were recruited through a variety of institutional mailing lists, social media, forums and participant recruitment websites, the aim being to recruit participants from a range of backgrounds. In total, 58 participants completed all three sessions of the experiment. 45 participants with valid email addresses completed the online questionnaire but did not complete either of the listening sessions and seven participants completed the first listening session but did not complete the second listening session. This resulted in a total attrition rate of 47%. It should be mentioned that this is a higher attrition rate than one might find in laboratory experiments. High attrition rates in web-based studies have previously been reported and discussed elsewhere (Mason and Suri, 2012). For more details about the participants, see Section 6.3.2.



(a) BIR mean difference per participant.

(b) BIR mean difference distribution.

Figure 6.2: BIR variance.

6.3 RESULTS

6.3.1 *Participant Reliability*

Prior to conducting data analysis on either the psychographic data or OLE data, participant suitability and reliability was assessed through several means. Firstly, two participants self-reported that they did not have normal hearing and were therefore excluded from the analysis.

In both listening sessions participants made basic item ratings of all 10 stereo items. To assess participant reliability it was therefore possible to calculate the mean rating difference between the basic item ratings in each session, see Figure 6.2. As seen in Figure 6.2b, the median BIR difference between the two sessions is 1, i.e. one star on the rating scale, and the distribution around the median is normal. Two participants are seen to have a BIR mean difference outside of 1.5 x the interquartile range, seen as outliers in Figure 6.2b. As these two outliers are close to the boundary of 1.5 x IQR (within 0.2 rating stars), it was decided that it was not necessary to exclude these participants from further analysis.

Finally, the distribution of each participants' BIRs were checked in order to identify participants who may skew the results. Participants with a mode BIR at the extremes of the rating scale (i.e. those who chose "not at all" or "very much" most frequently) would potentially be limited in expressing improvements or deterioration due to the processing in comparison to their BIRs, i.e. floor and ceiling effects, as further dis-

cussed in (Schoeffler and Herre, 2016). It could therefore be expected that by including such participants, the difference in OLE ratings with respect to the different quality levels would be reduced and also that any correlations relating item ratings to quality levels would be weakened. Eight participants had a mode BIR of either "not at all" or "very much" and were therefore excluded from further analysis. The relatively high number of participants excluded at this stage is likely a result of the split in content between classical and jazz items and more contemporary items, coupled with the wide range of backgrounds of the participants. An alternative approach to excluding participants would be to individually select the items presented to each participant as in previous OLE experiments (Schoeffler and Herre, 2016), although this requires a larger pool of items to be rated than available for this study.

To summarise, a total of 10 participants (17%) were excluded and therefore data from 48 participants are used in the following analysis.

6.3.2 *Psychographic Data*

In this section psychographic data from the remaining 48 participants are presented.

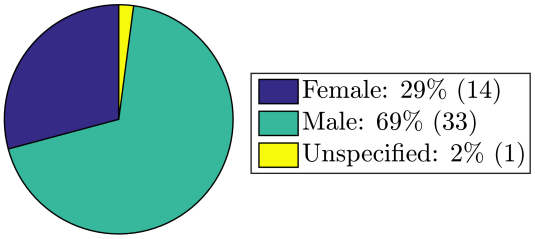
Demographics

Figure 6.3 shows the distribution of data from the psychographic questions relating to gender, age group and education level. It is seen that the sample is predominantly male (69%), younger than 35 (61%) and educated to a university level (73%).

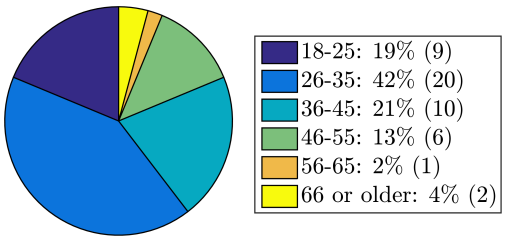
Experience

Figure 6.4 shows the distribution of data from the psychographic questions relating to audio technology in work and hobbies, headphone usage, binaural experience and previous listening tests. With regards to work and hobbies, the sample is equally split between those who have work and hobbies related to audio technology and those who do not. The majority of participants listen to audio over headphones either most days or often (77%)[1]. For binaural audio listening experience, half of the sample have

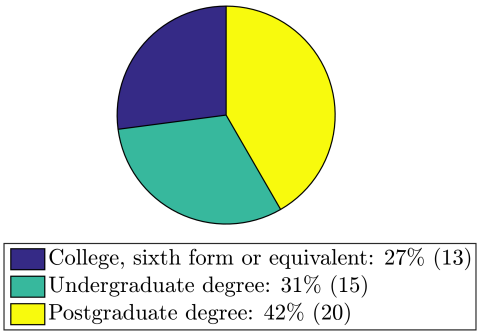
1 Comparing this value to other studies suggests that this is likely to be representative of a general population. For example, previous studies report figures of 57% for daily use of portable media players (PMPs) amongst Swedish adults (Kähäri, Åslund, and Olsson, 2011) (note that this is a more specific criterion than that used in this study) and 89% for daily or several times weekly use of PMPs amongst Swedish adolescents (Widén et al., 2017).



(a) Gender.



(b) Age group.



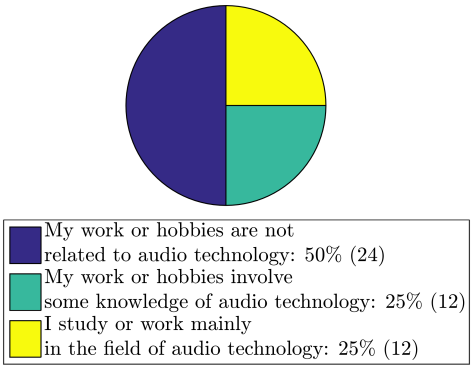
(c) Education.

Figure 6.3: Distribution of data from demographic related psychographic questions.

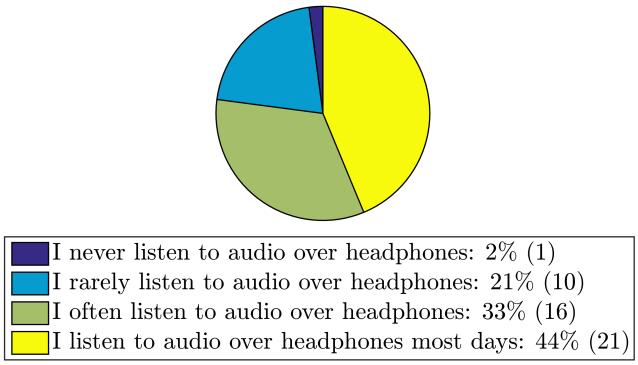
some experience of listening to binaural audio, 25% have no experience and 25% are not sure. It could be assumed that those who are not sure are unfamiliar with the term "binaural" and are therefore more likely to have no experience rather than some experience. If this is the case, the sample would be equally split between those who have no experience and those who have some experience of binaural audio listening. Finally, just less than half of the sample (46%) have not participated in listening tests previously. These distributions suggest that the participant sample is more audio technology minded than a general population. However, an approximately equal split in the sample regarding audio technology experience in general is beneficial for the aims of this experiment.

Attitudes Towards Audio Technology

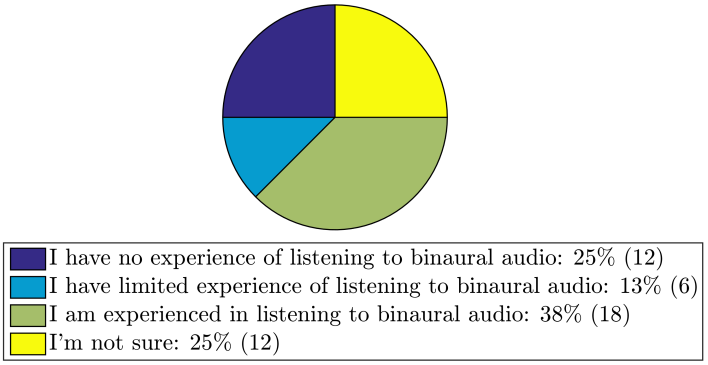
Normalized scores for the competence, enthusiasm and DSI scales were calculated by assigning values of 1-5 to the Likert scale responses, summing these values (including inverting values for negative questions) and normalizing by the number of questions each scale contained. This resulted in values for each participant between 0 and 1 for the three measures. Additionally, a combined "total" measure was created by taking the mean of each participants' competence, enthusiasm and DSI values. Figure 6.5



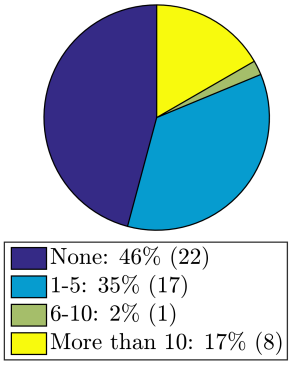
(a) Work and hobbies.



(b) Headphone usage.



(c) Binaural experience.



(d) Previous listening tests.

Figure 6.4: Distribution of data from experience related psychographic questions.

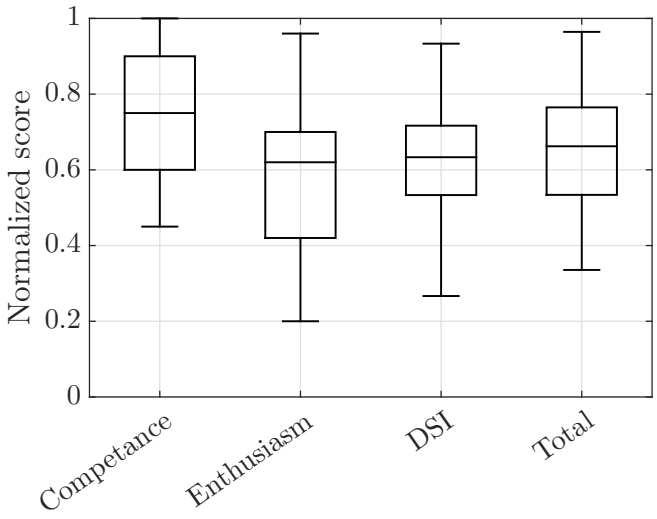


Figure 6.5: Distribution of data from attitude related psychographic questions.

shows the distribution of attitude values. It is seen that for all measures the median lies between 0.6 and 0.8, with competence having the highest median and enthusiasm the lowest. The smallest range in results is seen for competence (0.45 - 1) and the largest for enthusiasm (0.2 - 0.96). Despite the skew to higher scores, the diversity in attitudes is sufficient for the analysis presented in the following sections.

6.3.3 *OLE Analysis: Part I*

The OLE ratings were made on a five-star Likert scale and as such could either be interpreted as ordinal data (from the labels) or interval data (from the number of stars). Typically it is recommended to use non-parametric statistics and median values for ordinal data, whereas with interval data it is possible to use parametric statistics and mean values. The choice of analysis for Likert-type data is well discussed in the literature and some prominent studies such as (Norman, 2010) advocate the use of either non-parametric or parametric analysis. Specific to the analysis of OLE, it was shown that there are only minor differences in effect sizes and statistical significance values when comparing non-parametric and parametric methods (Schoeffler et al., 2015a). In the analysis of the OLE data presented here, the data are generally regarded as ordinal and as such non-parametric statistical techniques are used. In some cases mean values are deemed appropriate and are also included.

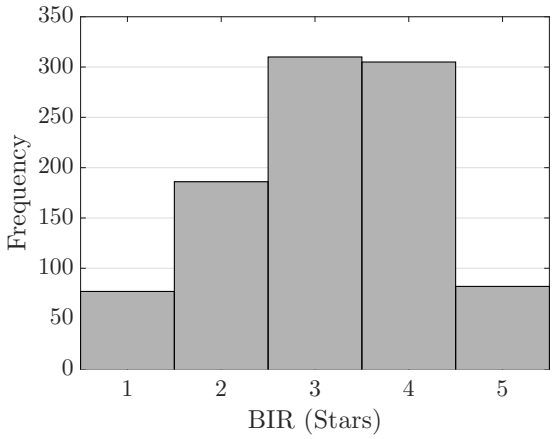


Figure 6.6: Histogram of all basic item ratings.

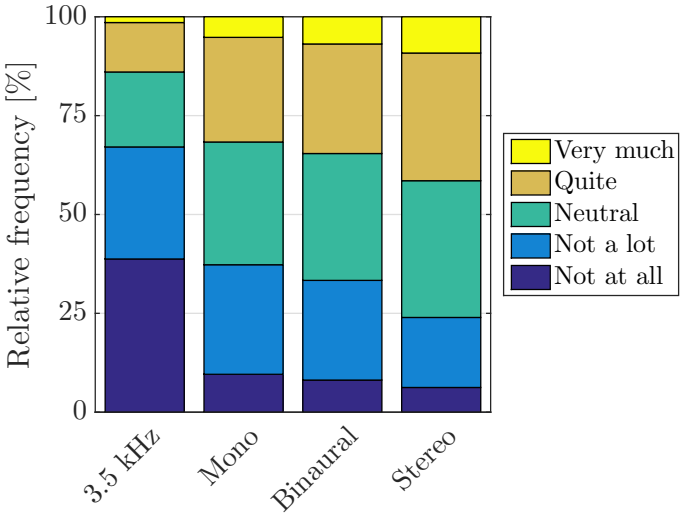


Figure 6.7: Relative frequencies of item ratings grouped by processing.

To gain an overview of how much the content was liked by participants, a histogram of all basic item ratings is presented in Figure 6.6. When averaging over participant and content, ratings of "neutral" and "quite" are most common followed by "not a lot". The relatively large difference between frequencies of the middle ratings (2, 3 and 4 stars) compared to the extreme ratings (1 and 5 stars) suggests that the content chosen is suitable to evaluate the impact of the various quality levels.

An overview of the OLE ratings associated with the different processing conditions can be seen in Figure 6.7. When averaged over the different items it is seen that the 3.5 kHz condition has the lowest ratings followed by mono, binaural and stereo. Non-parametric Wilcoxon signed-rank tests are used to quantify the significance of

Table 6.2: Z statistics and p-values from Wilcoxon signed-rank tests on OLE data.

	Mono	Binaural	Stereo
3.5 kHz	$Z = -12.5$ $p < .001$	$Z = -13.7$ $p < .001$	$Z = -14.5$ $p < .001$
Mono		$Z = -2.2$ $p = .030$	$Z = -6.8$ $p < .001$
Binaural			$Z = -4.8$ $p < .001$

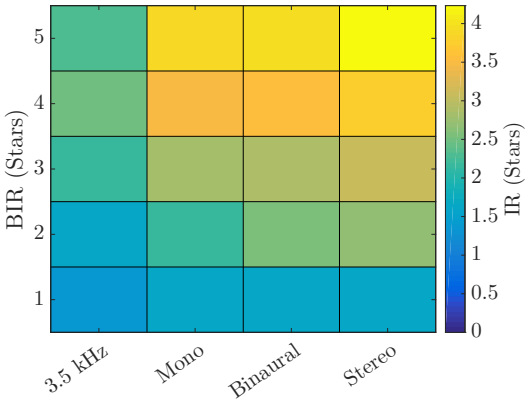


Figure 6.8: Colour map of average item ratings grouped by processing and basic item rating.

the differences between these conditions and values from this analysis are presented in Table 6.2. All comparisons reveal significant differences ($p < .05$). The timbral degradation introduced by a 3.5 kHz low-pass filter has much more of an impact on the ratings than either the mono or binaural processing. The small difference in ratings between mono, binaural and stereo suggest that, when averaged over participant and content, spatial processing has only a small affect on OLE. When comparing the stereo and binaural conditions, it is seen that binaural processing produces significantly lower ratings than stereo ($Z = -4.8$, $p < .001$), although the difference in ratings is small (an average of 0.2 stars).

Figure 6.8 presents a colour map of the OLE results grouped by processing and basic item rating. This shows how basic item ratings relate to item ratings for the different types of processing. From the figure it is seen again that the 3.5 kHz low-pass processing has the largest impact on ratings, with BIRs of five stars relating to an average IR of 2.3 stars for this type of processing. In comparison, the mono and binaural conditions both have IRs of 3.9 for five-star BIRs.

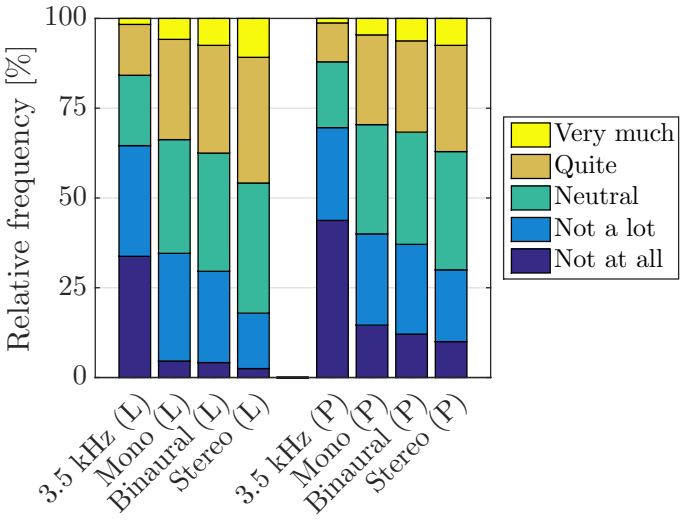


Figure 6.9: Relative frequencies of item ratings grouped by processing and content group. L=live, P=pop.

Due to the different characteristics and production techniques of the content, it is necessary to perform analysis of the OLE data with respect to content. Of particular interest is the comparison between the classical and jazz items, which were live concerts captured with a microphone array and spot microphones, and the remaining popular items, which were studio sessions captured with typical close microphone techniques only. As can be seen in Figure 6.9, the rating order with respect to processing is the same for each group. When comparing the binaural and stereo conditions, the difference in mean ratings between these is 0.25 stars for the live content and 0.16 stars for the pop content. This suggests that binaural processing had less of a negative impact on the pop items than the live items, although this difference is small.

Additionally, the OLE data was analysed with respect to each content individually. For seven of the 10 items the order of ratings with respect to processing was in line with the averaged results presented above. For two of the items (genres "Jazz trumpet" and "Pop", see Table 6.1) the mono condition was rated above the binaural condition. For one item ("Indie") the binaural and stereo conditions had equal mean ratings.

6.3.4 *Analysis of Listener Type*

In this section, the influence of quality and content on OLE is determined for each participant. As suggested by Schoeffler and Herre, (2014), this is achieved by calculating Kendall rank correlation coefficients (Kendall's τ). Kendall's τ is a non-parametric

statistic used to measure the ordinal association between two variables and results in a value ranging from -1 to $+1$. A value of -1 indicates perfect disagreement between the two variables, a value of 0 indicates that the two variables are independent and a value of $+1$ indicates perfect agreement between the two variables.

For each participant, four Kendall's τ values were calculated. To measure to what extent the content influences the OLE ratings, Kendall's τ was calculated from each participant's item ratings and basic item ratings ($\tau_{IR,BIR}$):

$$\tau_{IR,BIR} = cor_{\tau}(\mathbf{IR}, \mathbf{BIR}), \tag{5}$$

where **IR** and **BIR** are vectors of each participant's item ratings and basic items respectively. These vectors are sorted by item and processing and are therefore organised so that $\mathbf{IR}(i)$ and $\mathbf{BIR}(i)$ are ratings corresponding to the same item. To measure to what extent the timbral quality influences OLE ratings, Kendall's τ was calculated from each participant's item ratings associated with the 3.5 kHz and stereo conditions, and the associated timbral quality levels ($\tau_{IR,T}$):

$$\tau_{IR,T} = cor_{\tau}(\mathbf{IR}, \mathbf{T}), \tag{6}$$

where **T** is a vector containing the ranks of the timbral quality levels. The rank order of **T** is defined as: 3.5 kHz $<$ stereo. $\mathbf{T}(i)$ therefore identifies the timbral quality level of $\mathbf{IR}(i)$ as either 3.5 kHz or stereo. To measure to what extent the spatial quality influences OLE ratings, Kendall's τ was calculated from each participant's item ratings associated with the mono and stereo conditions, and the associated spatial quality levels ($\tau_{IR,S}$):

$$\tau_{IR,S} = cor_{\tau}(\mathbf{IR}, \mathbf{S}), \tag{7}$$

where **S** is a vector containing the ranks of the spatial quality levels. The rank order of **S** is defined as: mono $<$ stereo. $\mathbf{S}(i)$ therefore identifies the spatial quality level of $\mathbf{IR}(i)$ as either mono or stereo. It should be noted that the binaural condition is not included in the calculation of $\tau_{IR,S}$. One requirement for Kendall's τ analysis is that there is a monotonic relationship between the two variables. As such, it was decided to exclude the binaural quality level from the analysis as this quality level was not consistently rated between the mono and stereo quality levels when considering the results on a participant by participant basis. In other words, participants' ratings would not necessarily reflect the rank order of mono $<$ binaural $<$ stereo, thus breaking the assumption of a monotonic relationship between **IR** and **S**. Finally, to measure to what

extent the total quality influences OLE ratings, Kendall's τ was calculated from each participant's item ratings associated with the 3.5 kHz, mono and stereo conditions, and the associated quality levels ($\tau_{IR,Q}$):

$$\tau_{IR,Q} = cor_{\tau}(\mathbf{IR}, \mathbf{Q}), \tag{8}$$

where $\mathbf{Q}$ is a vector containing the ranks of the total quality levels. The rank order of $\mathbf{Q}$ is defined as: 3.5 kHz $<$ mono $<$ stereo. $\mathbf{S}(i)$ therefore identifies the total quality level of $\mathbf{IR}(i)$ as either 3.5 kHz, mono or stereo.

Figure 6.10 presents a scatter plot of $\tau_{IR,Q}$ values versus $\tau_{IR,BIR}$ values for each participant. In this plot (and the subsequent correlation plots), each data point represents correlation values associated with one participant. Furthermore, the marker type indicates whether each participant's correlation values are significant ($p < .05$) for the correlations in question, as calculated from the Kendall's τ analysis. For example, in Fig. 6.10, each participant's item ratings can be significantly correlated with the total quality level (red plus), their basic item ratings (black circle), neither total quality level nor BIRs (blue x), or both total quality level and BIRs (black circle filled with red plus), as determined by the Kendall's τ calculations of $\tau_{IR,Q}$ and $\tau_{IR,BIR}$. Those participants with a high $\tau_{IR,Q}$ value are heavily influenced by the technical audio quality when making OLE ratings and those participants with a high $\tau_{IR,BIR}$ value are heavily influenced by the content when making OLE ratings. Applying Pearson's correlation to the data reveals a strong negative correlation between the pairs of τ values ($r = -0.63$).[1] In other words, participants who are more influenced by technical audio quality are less influenced by content and *vice versa*. This is in line with results presented by Schoeffler and Herre, (2014), who reported that a continuum exists that describes to what extent a listener's OLE ratings are influenced by technical audio quality and content. From Fig. 6.10 it is also apparent that some listeners are weakly influenced by both quality and content, represented by the data points at low values on both axes.

To explore the relative impact of the timbral quality levels and the spatial quality levels on $\tau_{IR,Q}$, scatter plots of $\tau_{IR,T}$ versus $\tau_{IR,Q}$ and $\tau_{IR,S}$ versus $\tau_{IR,Q}$ are examined, Figure 6.11. Pearson's correlation reveals a very strong positive correlation ($r = 0.99$) between $\tau_{IR,T}$ and $\tau_{IR,Q}$ and a strong positive correlation ($r = 0.72$) between $\tau_{IR,S}$ and $\tau_{IR,Q}$. This shows that the total quality levels are much more influenced by the timbral

1 For effect sizes in behavioural research, Cohen's conventions are typically used (Cohen, 1988). These state that an r of $|.1|$ represents a 'small' effect size, $|.3|$ represents a 'medium' effect size and $|.5|$ represents a 'large' effect size.

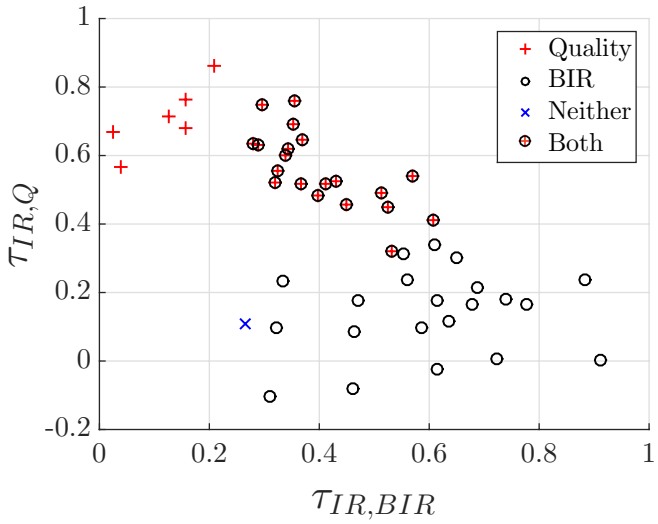


Figure 6.10: Kendall's rank correlations between item rating and the two variables total quality level ($\tau_{IR,Q}$) and basic item rating ($\tau_{IR,BIR}$) for each participant. Each data point represents correlation values associated with one participant. Marker type indicates significant correlations ($p < .05$) between item ratings and the factors indicated in the legend, as determined by the Kendall's τ analysis. 'Quality' refers to total quality level.

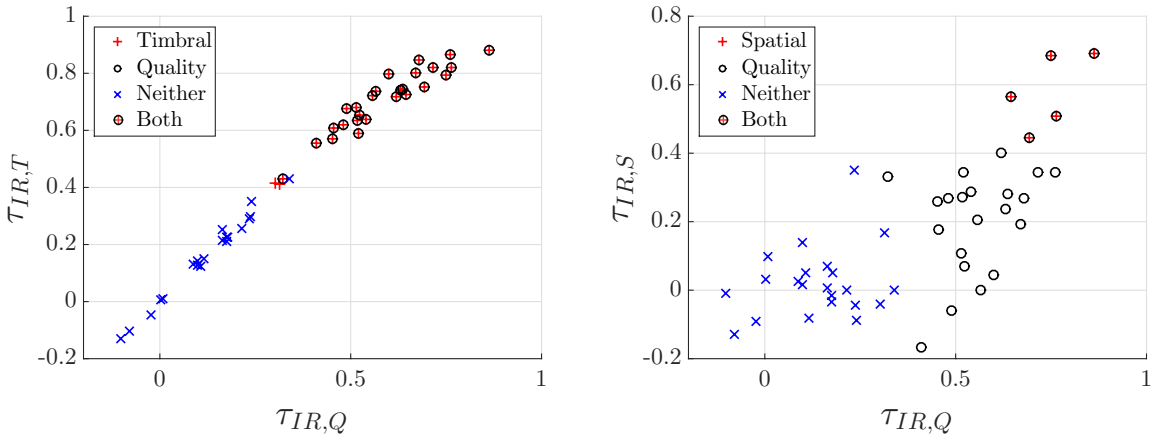


Figure 6.11: Kendall's rank correlations between item rating and the two variables timbral quality level ($\tau_{IR,T}$) and total quality level ($\tau_{IR,Q}$) (Figure a) and the two variables spatial quality level ($\tau_{IR,S}$) and total quality level ($\tau_{IR,Q}$) (Figure b). Each data point represents correlation values associated with one participant. Marker type indicates significant correlations ($p < .05$) between item ratings and the factors indicated in the legend, as determined by the Kendall's τ analysis. 'Timbral', 'Quality' and 'Spatial' refer to the timbral, total and spatial quality levels respectively.

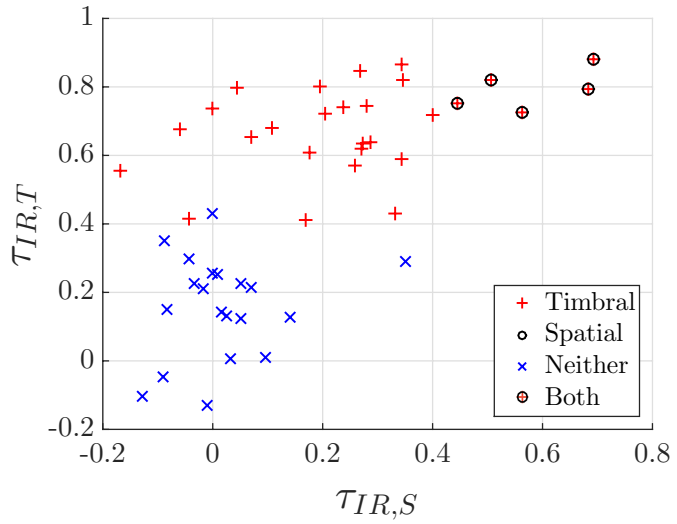


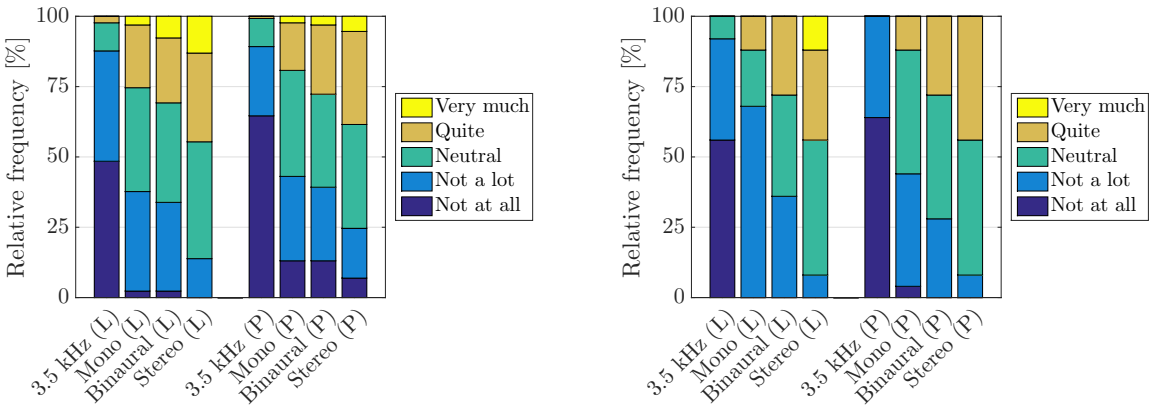
Figure 6.12: Kendall's rank correlations between item rating and the two variables timbral quality level ($\tau_{IR,T}$) and spatial quality level ($\tau_{IR,S}$) for each participant. Each data point represents correlation values associated with one participant. Marker type indicates significant correlations ($p < .05$) between item ratings and the factors indicated in the legend, as determined by the Kendall's τ analysis. 'Timbral' and 'Spatial' refer to timbral and spatial quality levels respectively.

levels than the spatial levels, and this is expected given the OLE ratings presented in Figure 6.7.

To assess if participants who are influenced by timbral quality are also influenced by spatial quality, $\tau_{IR,T}$ versus $\tau_{IR,S}$ is plotted, Figure 6.12. Pearson's correlation reveals a strong positive correlation ($r = 0.64$) which indeed suggests that participants who are influenced by timbal quality are more likely to be influenced by spatial quality. It is apparent from Figure 6.12 that only a small number of participants are significantly influenced by spatial quality (five) and all of these are also significantly influenced by timbral quality. Furthermore, there are some participants who are significantly influenced by timbral quality but have very low correlations with spatial quality. It could therefore be said that participants who are significantly influenced by spatial quality will typically be influenced by timbral quality, but this is not the case in reverse.

6.3.5 *OLE Analysis: Part II*

In Section 6.3.4 it was revealed that participants are influenced by technical audio quality by various amounts when making OLE ratings. This insight is now used to analyse the OLE ratings with respect to listener group.



(a) Ratings from 26 participants whose OLE is significantly influenced by total quality level.

(b) Ratings from five participants whose OLE is significantly influenced by spatial quality level.

Figure 6.13: Relative frequencies of item ratings grouped by processing, content and listener group.

Five participants' OLE ratings were significantly influenced by the spatial quality level and 26 participants' OLE ratings were significantly influenced by the total quality level. In order to evaluate whether these participants gave different OLE ratings compared to the average, relative frequency plots of OLE ratings are presented for each of these groups of listeners, Figure 6.13. As with the OLE ratings averaged over all participants, it is seen that both listener groups rate the processing in the order of stereo, binaural, mono and 3.5 kHz low-pass for both groups of content. The difference in mean rating between binaural and stereo is 0.46 stars for the group whose OLE is significantly influenced by spatial quality level, which is greater than the difference of 0.2 when using all participants. In terms of the comparison between the binaural and stereo conditions, it can therefore be concluded that the binaural processed items gave a lower OLE than the stereo items for those participants whose OLE is significantly influenced by spatial quality level, as well as for the participants as a whole.

6.3.6 *Interaction Between Psychographic Variables and Listener Type*

The interaction between the psychographic variables presented in Section 6.3.2 and the measures of listener type presented in Section 6.3.4 is now investigated. Except for gender, all of the psychographic variables are measured on an ordinal or continuous scale and as such Kendall's τ can be used to investigate correlations between

Table 6.3: Kendall's τ correlation and significance between psychographic variables and measures of listener type - $\tau_{IR,Q}$, $\tau_{IR,T}$, $\tau_{IR,S}$ and $\tau_{IR,BIR}$. *For gender, a point-biseiral correlation was used. Significant correlations are highlighted in red.

	$\tau_{IR,Q}$	$\tau_{IR,T}$	$\tau_{IR,S}$	$\tau_{IR,BIR}$
Gender*	$r = .487$, $p < .001$	$r = .484$, $p < .001$	$r = .335$, $p = .020$	$r = -.416$, $p = .003$
Age	$\tau = .078$, $p = .478$	$\tau = .080$, $p = .466$	$\tau = .128$, $p = .247$	$\tau = .006$, $p = .955$
Education	$\tau = .063$, $p = .581$	$\tau = .091$, $p = .424$	$\tau = -.061$, $p = .594$	$\tau = .059$, $p = .607$
Work / hobbies	$\tau = .442$, $p < .001$	$\tau = .454$, $p < .001$	$\tau = .270$, $p = .018$	$\tau = -.352$, $p = .002$
Headphones usage	$\tau = .207$, $p = .068$	$\tau = .199$, $p = .080$	$\tau = .165$, $p = .146$	$\tau = -.130$, $p = .251$
Binaural exp.	$\tau = .243$, $p = .028$	$\tau = .256$, $p = .021$	$\tau = .188$, $p = .092$	$\tau = -.181$, $p = .103$
Prev. listening tests	$\tau = .225$, $p = .049$	$\tau = .262$, $p = .022$	$\tau = .079$, $p = .487$	$\tau = -.265$, $p = .020$
Competence	$\tau = .537$, $p < .001$	$\tau = .549$, $p < .001$	$\tau = .302$, $p = .003$	$\tau = -.397$, $p < .001$
Enthusiasm	$\tau = .440$, $p < .001$	$\tau = .468$, $p < .001$	$\tau = .258$, $p = .012$	$\tau = -.324$, $p = .002$
DSI	$\tau = .322$, $p = .002$	$\tau = .353$, $p = .001$	$\tau = .145$, $p = .156$	$\tau = -.208$, $p = .042$
Total attitude	$\tau = .481$, $p < .001$	$\tau = .523$, $p < .001$	$\tau = .271$, $p = .007$	$\tau = -.346$, $p = .001$

the psychographic variables and measures of listener type (τ_Q, τ_T, τ_S and τ_{BIR}), see Table 6.3. As gender is a dichotomous variable, a point-biserial correlation is used instead. Spearman's rank correlation was also used to verify the significant correlations. It is seen that the variables age, education and headphone usage do not show any significant correlations with the measures of listener type. The remaining variables on the other hand, all show significant correlations with one measure of listener type or more. The variable competence shows the strongest correlation with the measures $\tau_{IR,Q}$ and $\tau_{IR,T}$. For $\tau_{IR,S}$ and $\tau_{IR,BIR}$, the strongest correlation is with gender. Care should be taken when interpreting this result however, and indeed the other gender correlations, as the sample was not equally stratified by gender. For example, with regards to work and hobbies no females answered "I study or work mainly in the field of audio technology" compared to 12 males. It therefore cannot be assumed that it is gender itself that leads to the significant correlations seen. Excluding gender, the strongest correlation with the measures $\tau_{IR,S}$ and $\tau_{IR,BIR}$ is also competence. As well as the variable competence, variables total attitude (which includes competence), work / hobbies and enthusiasm all show correlations above $\tau = 0.4$ for $\tau_{IR,Q}$.

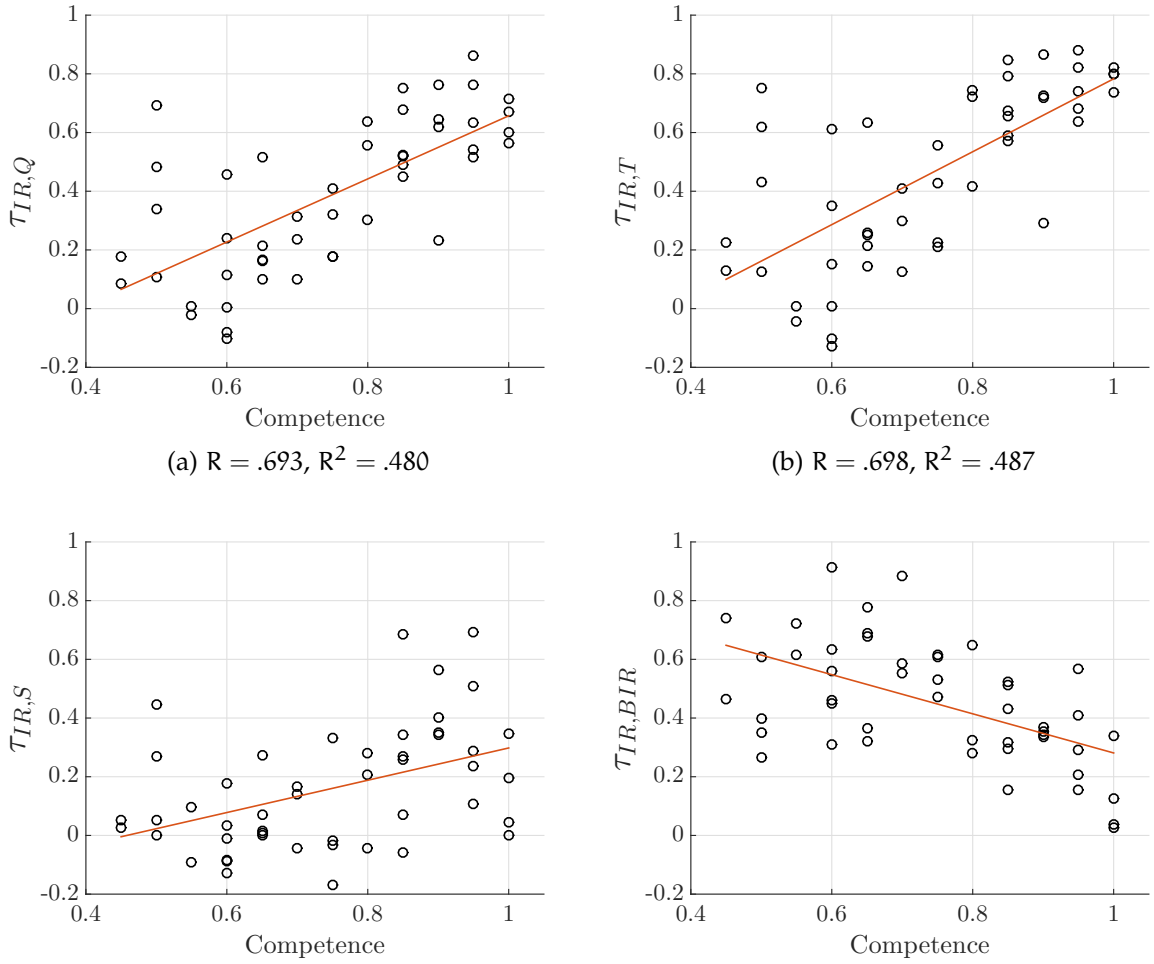
To predict measures of listener type from the psychographic variables, stepwise multiple regressions were performed for each measure. The independent variables in the regressions were the significant psychographic variables associated with each measure (presented in Table 6.3), excluding gender and also total attitude (due to possible multicollinearity problems with the variables that make up total attitude). The relevant assumptions related to multiple regression analysis were checked including

independence of residuals, linear relationships between the dependent and independent variables, homoscedasticity, multicollinearity issues and normal distribution of residuals.

For all four measures of listener type ($\tau_{IR,Q}$, $\tau_{IR,T}$, $\tau_{IR,S}$ and $\tau_{IR,BIR}$), competence was the only variable that added significantly to the prediction and, as such, all other variables were excluded from the model. The specific significance values and model coefficients are listed below.

- $\tau_{IR,Q} = -0.418 + (1.075 \times \text{competence})$
  $F(1,46) = 42.506, p < .0005, R = .693, R^2 = .480$
- $\tau_{IR,T} = -0.459 + (1.242 \times \text{competence})$
  $F(1,46) = 43.652, p < .0005, R = .698, R^2 = .487$
- $\tau_{IR,S} = -0.252 + (0.550 \times \text{competence})$
  $F(1,46) = 10.991, p = .002, R = .439, R^2 = .193$
- $\tau_{IR,BIR} = 0.948 - (0.667 \times \text{competence})$
  $F(1,46) = 19.293, p < .0005, R = .544, R^2 = .295$

From the above values, it is seen that competence explains 48% of the variance in $\tau_{IR,Q}$, 49% of the variance in $\tau_{IR,T}$, 19% of the variance in $\tau_{IR,S}$ and 30% of the variance in $\tau_{IR,BIR}$. The effect sizes (R values) can all be classified as strong, with the exception of the correlation between $\tau_{IR,S}$ and competence which can be classified as moderate. The correlations between the measures of listener type and competence are presented in graphical form in Figure 6.14. When looking at the plot of $\tau_{IR,Q}$ versus competence, it is seen that participants with a high competence score (> 0.8) have relatively similar $\tau_{IR,Q}$ values (within 0.4 of each other), with the exception of one participant. On the other hand, participants with a low competence score ($\leqslant 0.6$) show a larger range in $\tau_{IR,Q}$ values. That is to say, participants who have high competence scores are typically highly influenced by technical audio quality when making OLE ratings, although the opposite is less certain to be true for participants with low competence scores. On the other hand, when looking at the plot of $\tau_{IR,S}$ versus competence it is seen that the largest range of $\tau_{IR,S}$ values are for participants with high competence values. In other words, it is hardest to predict how much a listener will be influenced by spatial audio quality for participants who have high competence scores. Finally, when looking at the plot of $\tau_{IR,BIR}$ versus competence, a negative correlation is seen which shows that participants with high competence scores are typically less influenced by the content



(a) R = .693, R^2 = .480

(b) R = .698, R^2 = .487

(c) R = .439, R^2 = .193

(d) R = .544, R^2 = .295

Figure 6.14: Correlations between measures of listener type and the psychographic variable competence with regression lines plotted. Each point represents values from one participant.

than participants with low competence scores, and this is expected given the previous results.

6.3.7 *Complementary Analysis*

The analyses presented thus far have been predominantly based on correlation values, as calculated to identify listener type. To support the conclusions drawn in the above sections it is beneficial to also provide analysis based on the raw item ratings. Specifically, in this section we aim to support the above conclusion that the attitudinal measure 'competence' is a significant predictor of listener type by conducting an analysis of variance on the raw item ratings.

A three-way mixed ANOVA was conducted on normalized IR data with overall quality level (four levels) and content (10 levels) as within-subject factors, and competence (two levels) as a between-subject factor. Note that this is a parametric analysis and therefore the OLE ratings are considered as interval data, as discussed previously. The IR data was normalized to the BIR data by subtracting participants' BIRs from their IRs, where the BIRs are from the corresponding session. The normalized IR data therefore takes into account the degree of liking of the content and is a measure of the deviation between participants' basic item ratings and item ratings. To prepare the competence data for the ANOVA, it was necessary to transform it from continuous data to categorical data. This was achieved by dichotomising the competence scores into a low competence group (22 participants) and a high competence group (26 participants), split around the mean.

Prior to analysis the assumptions underlying the mixed ANOVA were checked, namely, normality for each combination of the within-subject and between-subject factors, homogeneity of variances for each combination of the groups and sphericity. The factor combinations were largely normally distributed with predominantly homogenous variances across between-subject factors. However, Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated for the within-subject factor of overall quality level ($\chi^2(5) = 29.1, p < .001$) and therefore a Greenhouse-Geisser correction was applied.

By studying the interaction between overall quality level and competence, it is possible to assess if the impact of overall quality on normalized item ratings (i.e. listener type) is influenced by competence. The interaction between overall quality level and competence was found to have a significant influence on the normalized item ratings

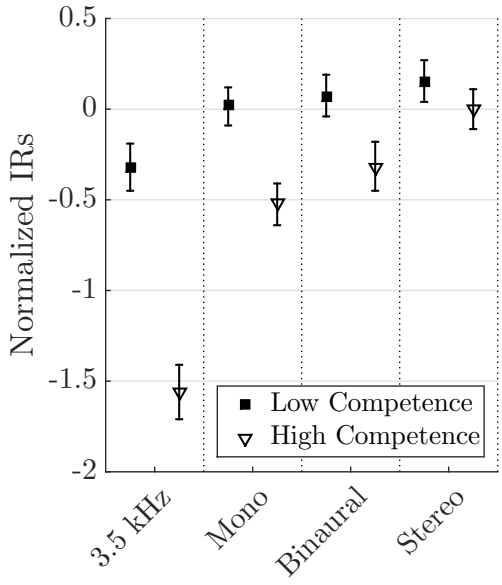


Figure 6.15: Normalized item ratings (averaged over content) with respect to both overall quality level and competence group. Error bars show 95% confidence intervals.

($F(2.20, 101) = 20.4, p < .001$). Furthermore, a partial eta-squared value of $\eta_p^2 = .307$ indicates a large effect. Figure 6.15 represents this interaction graphically, by plotting normalized item ratings (averaged over content) with respect to both overall quality level and competence group. It is seen that participants in the low competence group are only mildly influenced by the overall quality level. On the other hand, participants in the high competence group are significantly more influenced by the overall quality level, thus supporting the results from the previous sections.

6.4 DISCUSSION

The aims of this study were twofold. Primarily, correlations were sought between psychographic variables and the influence of technical audio quality on overall listening experience. Secondly, the influence of binaural audio on overall listening experience was investigated. The results regarding this second objective will be discussed first.

6.4.1 *OLE of Binaural Audio*

The influence of four processing conditions on OLE were compared; 3.5 kHz low-pass filter, mono, binaural and stereo. It was seen that all conditions have a statistically sig-

nificant influence on OLE. As expected, the 3.5 kHz low-pass filter conditions were rated as having the largest negative impact on OLE, with average ratings being 1.1 stars lower than for the stereo conditions. This result is roughly inline with previous studies (Schoeffler, Edler, and Herre, 2013). With regards to the spatially processed conditions (mono and binaural), the impact on OLE was seen to be much less pronounced. The ratings for the mono conditions were on average 0.3 stars lower than the stereo conditions and the ratings for the binaural conditions were on average 0.2 stars lower than the stereo conditions. For mono processing this is a slightly less pronounced influence on OLE than previously presented results (Schoeffler, Conrad, and Herre, 2014) where the difference was closer to 0.5 stars. The slight difference in the two studies may be due to the reproduction methods used as the previous research presented stimuli over loudspeakers. The stronger influence of timbral quality compared to spatial quality on listening experience is coherent with previous studies such as (Rumsey et al., 2005a), although a direct comparison cannot be made as total bandwidth between the timbral and spatial degradations were not matched in this study.

With regards to the specific research question of investigating the influence of binaural audio on OLE, it can be concluded that for the stimuli and participants used, binaural processing negatively influenced OLE in a small but significant manner. This was the case for both the live and studio groups of content, as well as for multiple groups of participants; those whose OLE ratings were significantly influenced by spatial audio quality, those whose OLE ratings were significantly influenced by total audio quality and the sample as a whole. As this study was purely quantitative in nature, it is not possible to say why the binaural content produced a lower OLE than the stereo content. One possible explanation could be that the binaural processing negatively influenced the timbral properties of the content. Further studies could help develop insight into this. It should be noted that despite the broad range of musical genres used in this study, the content was still limited to audio-only, music items. Further studies could investigate the OLE of binaural audio for other content types such as drama and audiovisual stimuli, where the possible spatial advantages offered by binaural processing could be more noticeable. Furthermore, the binaural processing used was non-personalised and static. Such binaural processing is common in broadcast binaural content and the evaluation of such processing is therefore valid. It would be interesting to see however, how results compare for binaural content with personalised HRTFs and head-tracking. Types of headphones used by participants could also

be an influencing factor on the results seen, although as with the type of processing, in reality binaural content is distributed to users with a range of headphone types.

6.4.2 *Human Factors and OLE*

The other aim of the study was to investigate correlations between psychographic variables and the influence of technical audio quality on overall listening experience. The first stage of this was to evaluate to what extent each participant was influenced by technical audio quality and the content when making OLE ratings. As with previous studies (Schoeffler and Herre, 2014), a negative correlation was found between the influence of quality and the influence of content on OLE ratings. Participants who are more influenced by technical audio quality are generally less influenced by content and *vice versa*. In previous studies labels of "audio quality likers" and "song likers" were used to describe this range in participants (Schoeffler and Herre, 2013). When looking at the influences of timbral audio quality and spatial audio quality on OLE, only five out of 48 participants were significantly influenced by spatial audio quality compared to 28 who were significantly influenced by timbral audio quality. All of those who were significantly influenced by spatial audio quality were significantly influenced by timbral audio quality. A strong positive correlation between the influence of timbral audio quality and the influence of spatial audio quality on OLE ratings was seen, which suggests that participants who are influenced by one aspect of quality are likely to be influenced by other aspects of quality.

Interactions between psychographic variables and listener type were studied by means of correlation and regression analysis. The psychographic variables that showed significant correlations with the influence of technical audio quality on OLE ratings included work and hobbies, binaural experience, previous listening tests, competence, enthusiasm, innovativeness and total attitude towards audio technology. To predict the influence of technical audio quality on OLE a multiple regression analysis was performed. The only psychographic variable that added significantly to the prediction was the attitudinal measure competence and indeed a strong correlation between competence and the influence of technical audio quality on OLE was seen ($R = .693$). This result suggests that, out of the psychographic variables studied in this experiment, the measure competence is the most useful for predicting to what extent a listener will be influenced by degradations in technical audio quality. This measure consists of four questions and is thus a practical way to quickly assess how a participant may respond

to different levels of technical audio quality. Applications of this knowledge could include adapting the technical audio quality requirements of a product or service to the potential users in a more educated way, improving quality prediction models and also using the competence questionnaire presented here for participant recruitment purposes in subjective evaluations. Typically in subjective evaluations of audio, data collected about participants includes professionalism and previous number of listening tests, although this study shows that gathering data about the attitudinal measure competence could be more worthwhile in some cases.

Despite there being a strong correlation between competence and influence of technical audio quality on OLE, a large variance around the regression line was seen. As this experiment was limited in the number of human influence factors studied, further studies should look for additional variables that help explain some of the variance not described by the variables used in this experiment. The method presented here would be suitable for further investigations and could be used to study factors such as emotions, mood, attention level and motivation. Moreover, the relative influence of content and technical audio quality on OLE is likely to be context specific as well as user specific. It would therefore be worthwhile investigating how these groups of influence factors interrelate with regards to OLE. For instance, it may be the case that some participants are heavily influenced by technical audio quality in a home context, but not in a mobile listening context. As the method used in this study is relatively simple and time-efficient to conduct, it would be possible to administer such a study on mobile devices in a range of contexts and environments to investigate such relationships.

6.5 SUMMARY

In this chapter, the role of human influence factors on the QoE of next generation audio were investigated. Relationships between a range of psychographic variables and the influence of technical audio quality on overall listening experience were studied, as well as the overall listening experience of binaural audio. With regards to the specific objectives, it was seen that listener type is significantly correlated with multiple psychographic variables and that the attitudinal measure 'competence' is the most suitable variable to be used as a predictor of listener type. Furthermore, it was seen that binaural processing negatively influenced OLE in a small but significant manner for the stimuli used. The results presented highlight the importance of considering

human IFs when designing and evaluating next generation audio. The method used here to investigate human IFs in relation to OLE was identified as being suitable for further studies probing the role of human IFs on the QoE of next generation audio.

7

DISCUSSION AND CONCLUSIONS

The work presented in this thesis has explored the subject of the quality of experience of next generation audio. At the beginning of this thesis, two aims were outlined for this research in the scope of this topic. One aim was to explore the role of three classes of influence factor, representing the influence of system, context and human, on the quality of experience of next generation audio. Secondly, it was an aim to investigate suitable methods and approaches for the evaluation of the quality of experience of next generation audio, with respect to its various influence factors. These aims were achieved by conducting five experimental studies covering three case studies, related to system, context and human IFs respectively.

In this chapter, we will summarise and discuss the research presented in the thesis thus far. After a brief reflection on the current state of the field by considering the literature review, the empirical studies are summarised and discussed in relation to the overall aims of the thesis. The contributions of the thesis as a whole are also considered before discussions are presented on possible further work in this area of research.

7.1 THE LITERATURE

The literature review presented in chapters 2 and 3 set the background for this thesis. It was seen that next generation audio can be characterised by emerging technological trends, such as immersive multichannel reproduction and object-based technologies, and by the associated immersive, interactive and personalised experiences they provide. Multichannel loudspeaker setups and binaural audio are at the forefront of delivering immersive experiences to consumers. Systems that may be more practical in domestic environments however, such as soundbars and the approach of media device orchestration, will likely play a more prominent role in immersive audio reproduction in the years to come. Object-based audio is playing a role in enabling immersive reproduction by allowing for the adaptation of content to better suit the device. It also offers potential benefits such as adapting content to suit the user and the environment.

An important aspect of developing new audio technology is evaluating its perceived quality. By studying the quality-formation process, we saw that perceived quality is multivariate in nature, is influenced by temporal and contextual factors and is unique to the individual. Quality of experience is a measure that encompasses these features and offers a user-centric approach to quality evaluation. It was seen that other research fields, such as HCI and UX, are also based on a user-centric mindset, although in the case of audio evaluation a more technology-centric approach is traditionally used. It was proposed that in order to assess and improve the experience provided by next generation audio, a QoE mindset should be taken that considers its various influence factors - system, context and human.

With regards to previous work, before commencing this research project the QoE of next generation audio was still a relatively unexplored area. System influence factors are a common area of study in the field of audio quality evaluation although this is not the case for context and human influence factors. Typically for the assessment of system influence factors, standardised methods are used that are more representative of technology-centric approaches than user-centric QoE approaches. It was therefore apparent from the literature review that there is a need to investigate system IFs of next generation audio with a QoE mindset, as well as a need to investigate the role of context and human IFs on the QoE of next generation audio. It is with this in mind that the aims of the thesis were formed.

7.2 PART I: SYSTEM

7.2.1 *Summary*

The specific objective of the first study was to subjectively compare soundbar technology with discrete surround technology. This was a valuable objective as soundbar technology may play an important role in delivering immersive audio to consumers in the coming years. Prior to this research, the perceptual evaluation of such technology was however very limited.

To achieve this objective it was decided to use the method open profiling of quality, a perceptual evaluation method developed in the field of multimedia QoE evaluation that combines preference ratings with sensory profiling. Summarising the results, it was found that preference ratings for the two soundbars were significantly lower than the discrete surround system and the discrete stereo system due to a combination of

both timbral and spatial factors. Participants' preferences were mapped to wide, enveloping and immersive items, which correlated to the discrete system. With regards to the method, the modified OPQ method was shown to be a valuable tool for developing insight into the interplay between reproduction system, experienced quality features and overall experience for a range of listeners.

7.2.2 *Discussion*

The results presented are of course limited in the fact that only a small sample of soundbars were studied with a limited range of content items. Despite this, the specific results from this study provide a valuable contribution as they indicate the general current state of soundbar technology. Prior to this study there was little formal understanding of the experience provided by soundbar technology, with only one limited study existing (Moulin, Nicol, and Gros, 2012).

Referring back to the overall aims of the thesis, it is clear that the system IF of reproduction system can have a large influence on QoE, and indeed this is nothing new; system IFs are widely studied in the field of audio quality evaluation. This study has however been more insightful with regards to the aim of investigating suitable methods for the evaluation of the QoE of next generation audio with respect to its various influence factors. The adapted OPQ method proved a valuable way to relate preference ratings with relevant quality features and allowed listeners to efficiently and effectively communicate their perceived listening experience. The method could therefore be used for further studies investigating the role of system IFs on QoE. For instance, the method would be particularly well suited to evaluate technology that can potentially deliver novel experiences, such as media device orchestration (Francombe et al., 2017a) and other innovative immersive technologies. The insight gained from the qualitative aspect of this method supports the use of mixed methods for understanding multimodal quality perception (Strohmeier, Jumisko-Pyykkö, and Kunze, 2010). Therefore, to fully understand the influence of system IFs on the QoE provided by next generation audio technology, quantitative methods such as those outlined by the ITU (ITU-R, 2015c; ITU-T, 1996) are of limited use on their own.

As we know, QoE is mutually influenced by system, context and human IFs. As well as system IFs, in this study human IFs were also considered by comparing preference ratings and elicited attributes from naïve and experienced listener groups. Due to the small sample size, differences between the groups of listeners were somewhat

inconclusive. It is possible however that with a larger sample size more significant differences would become apparent. In terms of context IFs, it is likely that a range of factors could influence the perceived QoE of soundbar technology, not least the acoustics of the room. To gain a deeper understanding of the experience provided by soundbar technology, context IFs should be addressed more thoroughly in future studies.

7.3 PART II: CONTEXT

7.3.1 *Summary*

The specific objective of the studies related to context IFs was to investigate whether environmental noise influences preferred audio object balance. This was a worthwhile objective as object-based audio offers the potential to adapt content to better suit the environment. However, no previous studies had empirically investigated how content should be adapted in noisy environments to provide a better QoE.

To achieve this objective three studies were conducted covering a range of approaches; both laboratory-based and web-based methods were used including both qualitative and quantitative aspects. The results across the three studies showed that environmental noise can significantly influence preferred BG-FG audio object balance and that the nature of preferences are very much dependent upon the individual. Thematic analysis showed that some participants chose to increase the background components in order to mask the environmental noise and to ensure that the background components were audible above the noise. Other participants chose to increase the foreground components in order to improve the speech intelligibility and also the comfort of the overall experience.

7.3.2 *Discussion*

The main contribution of these studies is that they empirically show how object-based content adaptation can be utilised to improve the listening experience in relation to environmental factors. This idea had previously been discussed in the literature, for example by Parmentier, (2015), although there was no empirical understanding of what content adaptations may benefit the listener and why. Previous studies to improve the listening experience in noisy environments focussed on modifying the content as a

whole (Mason et al., 2015; Reis, Carriço, and Duarte, 2009). This study builds on this previous work by exploring the benefits of object-based audio and hopefully leads the way for more such studies. It is likely that the results are heavily dependent upon the nature of the audio content and environmental noise. One limitation of this set of studies is the limited range of content used and therefore the effect of content and content-noise interaction on audio object balance needs to be investigated further.

Referring back to the overall aims of the thesis, these studies have shown that context IFs, namely environmental noise, are indeed relevant when evaluating the quality provided by next generation audio. By considering context IFs, the QoE provided by next generation audio can be improved and object-based audio is one application where this consideration is particularly relevant. Compared to system IFs, the methods used to study context IFs may be more dependent upon the specific IF being investigated. Replicating the relevant context IFs in a laboratory-based scenario is useful although this may not be applicable for all types of context IFs, such as some physical and social context IFs. The alternative approach is in-the-wild testing, in which case it is important to understand and report the various factors that make up a given context, for example by using the method proposed by Jumisko-Pyykkö and Utriainen, (2011).

As well as the context IF of environmental noise, in the first of the three studies interactions were sought between reproduction methods (stereo versus binaural), environmental noise and preferred audio object balance. Although no significant interactions were found, interactions between context and system IFs when evaluating preferred audio object balance should not be ruled out in future studies. The individual nature of the results highlight probable interactions between context and human IFs. By studying variables related to human IFs in future studies, it may be possible to predict how a listener would adjust content to their liking in noisy situations. Other context IFs, such as the task context, are also likely to be relevant for the optimisation of audio object balance in a range of contexts and could be the subject of future studies.

7.4 PART III: HUMAN

7.4.1 *Summary*

The primary specific objective of the final study was to investigate the role of human influence factors on overall listening experience. Previous research had shown that the relative influence of content and technical quality on OLE depends on the individual; on the one hand some users are heavily influenced by content when making OLE judgements and, on the other hand, some users are heavily influenced by technical audio quality when making OLE judgements, with a continuum of users between (Schoeffler and Herre, 2014). However, it was not known if listener type could be characterised, or even predicted, from variables related to the attitudes and demographics of the listeners. As it would be beneficial to be able to tailor content to the individual, investigating the role of human influence factors on overall listening experience was a worthwhile objective.

To achieve this a web-based study was conducted whereby participants first completed a questionnaire. This questionnaire collected data about participants' demographics, experience and attitudes towards audio technology. Participants then completed an online listening task based on the OLE methodology, whereby the overall listening experience of various quality levels of content was assessed. From this data it was possible to identify to what extent listeners were influenced by content and technical audio quality when making OLE ratings and then relate this to the various psychographic variables.

Results showed that a range of psychographic variables were significantly correlated with the influence of technical audio quality on OLE, including work and hobbies, binaural experience, previous listening tests, competence, enthusiasm, innovativeness and total attitude towards audio technology. Out of the psychographic variables studied it was the attitudinal measure "competence" that showed the strongest correlation with the influence of technical audio quality on OLE. Another objective of this study was to evaluate the influence of binaural processing on OLE, as this had not previously been investigated. Results concerning this objective showed that, for the content used, binaural processing influenced OLE in a small but negative manner.

7.4.2 *Discussion*

This study has built on the previous research by Schoeffler and Herre, (2014) to provide the valuable contribution of relating human factors to the relative influence of content and technical quality on overall listening experience. Applications of these results could include adapting the technical audio quality requirements of a product or service to potential users in a more educated way, improving quality prediction models and also using the competence questionnaire presented for participant recruitment purposes in subjective evaluations. Furthermore, this study contributes to the limited research on human influence factors in audio quality evaluation in general.

Referring back to the overall aims of the thesis, this experiment has shown that human IFs can play an important role when evaluating the quality provided by next generation audio, and that by studying human IFs, a greater understanding of QoE can be achieved. As one benefit of next generation audio is personalisation, it is important to understand how different users should be provided for with regards to content adaptation. The method presented here of an in-depth psychographic questionnaire followed by ratings to assess the influence of different factors on OLE, could be used to study a range of other human IFs, such as emotions, mood and motivation, as well as other system factors. The method of OLE is also useful as a stand alone method to evaluate the overall experience provided by next generation audio. However, compared to methods that include sensory profiling such as OPQ, the insight gained is limited as it is not possible to say which quality features are related to the given ratings.

Due to the way in which this experiment was designed, human IFs were studied in relation to system IFs. However, it is also likely that human IFs are linked to context IFs. For instance, some participants may be heavily influenced by audio quality in a home context but not in a mobile listening context. This interplay of influence factors may be especially important when studying individual preferences for next generation audio technology and should be considered in future studies.

7.5 OVERALL CONTRIBUTION

By considering three different case studies relating to the key technological trends and associated experiences of next generation audio, the roles of system, context and human IFs on the quality of experience of next generation audio have been explored.

In the field of audio quality evaluation, context and human IFs are often overlooked, yet for the case of next generation audio, they will likely play important roles in the improvement of QoE. This has been illustrated by this research. On a high-level, this work therefore contributes the thesis that to effectively evaluate the perceived quality of next generation audio, a QoE mindset should be taken that considers system, context and human influence factors. It should be mentioned that in this project different case studies were chosen so as to highlight the roles of the various classes of IFs. However, if a deep understanding of a particular technology is required, it would be beneficial to design a series of studies considering these influence factors solely for the technology under study. For instance, if one wanted to study the QoE provided by soundbars in-depth, it would be necessary to evaluate the perceived quality not just in an ideal environment with experienced assessors, but also with respect to various context and human IFs.

A range of other research has previously discussed QoE in relation to next generation audio assessment, for example (Nicol et al., 2014) and (Schoeffler, Silzle, and Herre, 2017), and therefore the link between QoE and audio evaluation is, by itself, not a novel contribution. However, this thesis furthers previous work linking QoE and next generation audio evaluation by explicitly discussing and investigating the various classes of QoE influence factors in relation to the experience provided by next generation audio. As seen earlier in the thesis, studies and discussions on QoE influence factors have also previously been made in relation to other areas of multimedia quality research. For example, over a range of studies Jumisko-Pyykkö, (2011) explored system, context and human/user IFs in relation to the evaluation of mobile television. This thesis supports the benefits of considering such influence factors and also plays a part in ensuring that audio quality research is in line with relevant findings and approaches from other domains of quality research.

Furthermore, this thesis contributes insights into appropriate methodological approaches to study the range of influence factors. Specific methodological contributions were highlighted above, so here methodological insights are discussed in a more general sense. In terms of quantification of impression, a characteristic of the approaches used in this thesis is the quantification of the terms “preference” or “enjoyment”, as opposed to the more typically used term “quality”. Moreover, several of the studies presented in this thesis utilised qualitative approaches. Both of these aspects are related to how QoE is defined. The definition of QoE refers to “the degree of delight...” (Möller and Raake, 2014), illustrating the affective nature of QoE, and the user de-

pendent nature of QoE means that qualitative approaches are often necessary to fully understand the provided experience. Affective and mixed methods approaches have of course been used before for QoE evaluation, such as by Jumisko-Pyykkö, (2011), Strohmeier, (2011) and Schoeffler, (2017). The successful use of such approaches in this thesis is still beneficial however, as it supports their use for the specific case of the assessment of the QoE of next generation audio.

Another methodological consideration to discuss is the benefits of laboratory-based versus web-based approaches, as in the studies presented, both laboratory-based and web-based approaches were used. Unlike for traditional quality evaluation procedures such as (ITU-R, 2015c; ITU-T, 1996), laboratory-based approaches are not always necessary for QoE assessment. As an alternative, web-based approaches can be used to improve the external validity of an experiment whilst increasing the sample size. With web-based approaches participants are more likely to be in an environment and mindset that is more representative of real-world situations compared to laboratory-based experiments. This could be especially useful when studying certain context and human IFs. There are of course limitations associated with web-based studies, namely lack of control over environmental and system related variables. By appropriately instructing participants, these limitations can be sufficiently managed for many QoE related studies. It could also be possible to record data about these variables, such as environmental noise data from device microphones, to ensure some degree of internal validity in a more robust manner. Studies in this thesis have shown how web-based approaches might be used to study QoE influence factors in relation to next generation audio. For future studies on the QoE of audio, informed decisions should be made about what approach is most suitable by weighing up the above advantages and limitations.

As well as addressing the aims and objectives set out at the start of this thesis, this research has produced new questions to be answered and has helped pave the way for future research on the quality of experience of next generation audio in general. These areas of further work are discussed in the following section.

7.6 FURTHER WORK

Areas of possible further work in relation to the specific objectives have previously been discussed in the relevant chapters. In this section, a few examples of areas of further work in relation to the overall topic of the thesis shall be discussed.

As was mentioned in Section 2.2.4, one trend of next generation audio technology is towards more practical approaches for delivering immersive audio in domestic environments, for instance with technology such as soundbars and the approach of media device orchestration. For such technology, it would be extremely beneficial to evaluate them in their respective contexts of use due to the possible influence of context IFs outlined in this thesis. To achieve this, researchers in the field of audio quality evaluation should look towards research fields such as UX, HCI and QoE, where in-the-wild testing is more commonplace. Furthermore, in-the-wild testing would be appropriate for headphone reproduction technology, whereby context IFs may play a prominent role in the perceived experience. For instance, it would be beneficial to repeat the studies presented in both Part I and Part II of this thesis as in-the-wild studies, to compare with the laboratory and web-based results.

In terms of human IFs, there are still a range of influence factors that need to be explored in relation to next generation audio. Emotions, mood, personality, motivation, attention, expectations and needs could all be studied in relation to object-based audio, as well as in relation to quality perception of audio in general. Physiological measures, as discussed in Section 3.5.1, could prove especially useful when examining such IFs.

Another area of further work could be to apply additional quality evaluation methods from the field of QoE to the evaluation of next generation audio. For example, a quality evaluation method was presented by Robitza, Garcia, and Raake, (2015) in which quality was evaluated without the need to repeat stimuli, so that the experiment was more enjoyable and externally valid compared to more traditional methods. Such methods could be combined with in-the-wild approaches to provide realistic consumption scenarios and thus more externally valid results. Other methods that could be applied to the quality assessment of next generation audio could include those that use physiological measures and web-based approaches that include crowdsourcing aspects.

7.7 CONCLUDING REMARKS

We began this thesis by remarking that the evolution of audio reproduction technology provides for more meaningful experiences to its users. It is hoped that the knowledge gained from this thesis will play some role in this evolution. Effectively evaluating the experience provided by next generation audio technology is an important aspect

of designing the technology so as to provide more meaningful experiences. Taking a QoE mindset that considers system, context and human influence factors is one way to achieve this.

REFERENCES

AES (2015). *AES TD*1004.1.15-10 *Recommendation for loudness of audio streaming and network file playback*. Audio Engineering Society Technical Document.

Agnew, P. W. and Kellerman, A. S. (2008). "Fundamentals of multimedia." In: *Multimedia Technologies: Concepts, Methodologies, Tools, and Applications*. Ed. by Rahman, S. M. IGI Global.

Armstrong, M., Melchior, F., Churnside, A., Shotton, M., Brooks, M., Evans, M., and Melchior, F. (2014). "Object-based broadcasting - Curation, responsiveness and user experience." In: *International Broadcasting Convention (IBC) 2014 Conference*.

Arndt, S., Antons, J. N., Schleicher, R., Möller, S., and Curio, G. (2012). "Perception of low-quality videos analyzed by means of electroencephalography." In: *4th International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 284–289.

BBC R&D, Jasmine Cox (2013). *Object-based broadcasting*. Online: http://www.bbc.co.uk/rd/blog/2013-05-object-based-approach-to-broadcasting. Accessed: 2017-09-29.

BBC R&D, Tom Parnell (2017). *Binaural audio at the BBC Proms*. Online: http://www.bbc.co.uk/rd/blog/2016-09-binaural-proms. Accessed: 2017-09-29.

Bech, S. (1992). "Selection and training of subjects for listening tests on sound-reproducing equipment." In: *Journal of the Audio Engineering Society* 40.7/8, pp. 590–610.

Bech, S. (1999). "Methods for subjective evaluation of spatial characteristics of sound." In: *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*.

Bech, S. and Zacharov, N. (2006). *Perceptual audio evaluation - Theory, method and application*. John Wiley & Sons, Ltd.

Bech, S., Hamberg, R., Nijenhuis, M., Teunissen, K., Looren de Jong, H., Houben, P., and Pramanik, S. K. (1996). "Rapid perceptual image description (RaPID) method." In: *Proc. SPIE 2657, Human Vision and Electronic Imaging*, pp. 317–328.

Begault, D. R., Wenzel, E. M., and Anderson, M. R. (2001). "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source." In: *Journal of the Audio Engineering Society* 49.10, pp. 904–916.

Beresford, K., Ford, N., Rumsey, F., and Zielinski, S. K. (2006a). "Contextual effects on sound quality judgements: Listening room and automotive environments." In: *Audio Engineering Society Convention 120*.

Beresford, K., Ford, N., Rumsey, F., and Zielinski, S. K. (2006b). "Contextual effects on sound quality judgements: Part II - Multi-stimulus vs. single stimulus method." In: *Audio Engineering Society Convention 121*.

Berg, J. (2005). "OPAQUE - A tool for the elicitation and grading of audio quality attributes." In: *Audio Engineering Society Convention 118* 4, pp. 1665–1672.

Berg, J. (2006). "How do we determine the attribute scales and questions that we should ask of subjects when evaluating spatial audio quality?" In: *Spatial Audio & Sensory Evaluation Techniques*. Guildford, UK.

Berg, J. (2009). "The contrasting and conflicting definitions of envelopment." In: *Audio Engineering Society Convention 126*.

Berg, J. and Rumsey, F. (1999). "Spatial attribute identification and scaling by repertory grid technique and other methods." In: *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*.

Berg, J. and Rumsey, F. (2000). "In search of the spatial dimensions of reproduced sound: Verbal protocol analysis and cluster analysis of scaled verbal descriptors." In: *Audio Engineering Society Convention 108*.

Berg, J. and Rumsey, F. (2002). "Validity of selected spatial attributes in the evaluation of 5-channel microphone techniques." In: *Audio Engineering Society Convention 112*.

Berg, J. and Rumsey, F. (2003). "Systematic evaluation of perceived spatial quality." In: *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*.

Berkhout, A. J., Vries, D. de, and Vogel, P. (1993). "Acoustic control by wave field synthesis." In: *The Journal of the Acoustical Society of America* 93.5, pp. 2764–2778.

Blauert, J. (2005). "Analysis and synthesis of auditory scenes." In: *Communication Acoustics*. Ed. by Blauert, J. Springer Berlin Heidelberg, pp. 1–25.

Blauert, J. and Jekosch, U. (2003). "Concepts behind sound quality: Some basic considerations." In: *Proceedings of Inter-Noise 2003*, pp. 72–76.

Blauert, J. and Jekosch, U. (2012). "A layer model of sound quality." In: *Journal of the Audio Engineering Society* 60.1/2, pp. 4–12.

Blumlein, A. (1931). *Improvements in and relating to sound transmission, sound recording and sound reproducing systems*. British Patent Specification 394325.

Braun, V and Clarke, V (2006). "Using thematic analysis in psychology." In: *Qualitative Research in Psychology* 3.2, pp. 77–101.

Breinbauer, H. A., Anabalón, J. L., Gutierrez, D., Cárcamo, R., Olivares, C., and Caro, J. (2012). "Output capabilities of personal music players and assessment of preferred listening levels of test subjects: Outlining recommendations for preventing music-induced hearing loss." In: *Laryngoscope* 122.11, pp. 2549–2556.

Cecchi, S., Virgulti, M., Primavera, A., Piazza, F., Bettarelli, F., and Li, J. (2016). "Investigation on audio algorithms architecture for stereo portable devices." In: *Journal of the Audio Engineering Society* 64.1/2, pp. 75–88.

Choisel, S. and Wickelmaier, F. (2006). "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound." In: *Journal of the Audio Engineering Society* 54.9, pp. 815–826.

Choisel, S. and Wickelmaier, F. (2007). "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference." In: *Journal of the Acoustical Society of America* 121.1, pp. 388–400.

Churchill, E. F. and Bardzell, J. (2007). "From HCI to media experience: Methodological implications." In: *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not As We Know It - Volume 2*. BCS-HCI '07, pp. 213–214.

Churnside, A. (2016). "Object-based radio: Effects on production and audience experience." PhD thesis. University of Salford.

Cobos, M., Lopez, J. J., Navarro, J. M., and Ramos, G. (2015). "Subjective quality assessment of multichannel audio accompanied with video in representative broadcasting genres." In: *Multimedia Systems* 21.4, pp. 363–379.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd edition)*. Routledge.

Conetta, R., Rumsey, F., Zielinski, S., George, S., Jackson, P. J. B., Dewhirst, M., Bech, S., and Meares, D. (2008). "QESTRAL (part 2): Calibrating the QESTRAL model using listening test data." In: *Audio Engineering Society Convention 125*.

Czyzewski, A., Ciarkowski, A., Kostek, B., Kotus, J., Lopatka, K., and Suchomski, P. (2016). "Adaptive personal tuning of sound in mobile computers." In: *Journal of the Audio Engineering Society* 64.6, pp. 405–428.

Davies, W. J., Bruce, N. S., and Murphy, J. E. (2014). "Soundscape reproduction and synthesis." In: *Acta Acustica united with Acustica* 100.2, pp. 285–292.

Dewhirst, M., Conetta, R., Rumsey, F., Jackson, P. S. B., Zielinski, S., George, S., Bech, S., and Meares, D. (2008). "QESTRAL (Part 4): Test signals, combining metrics and the prediction of overall spatial quality." In: *Audio Engineering Society Convention 125*.

Dijksterhuis, G. (1996). "Procrustes analysis in sensory research." In: *Multivariate Analysis of Data in Sensory Science*. Ed. by Naes, T. and Risvik, E. Elsevier.

Drossos, K., Mimilakis, S. I., Floros, A., and Kanellopoulos, N. (2012). "Stereo goes mobile: Spatial enhancement for short-distance loudspeaker setups." In: *Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*.

EBU (2014). *EBU R128 - Loudness normalisation and permitted maximum level of audio signals*. European Broadcasting Union.

Ebem, D. U., Beerends, J. G., Van Vugt, J., Schmidmer, C., Kooij, R. E., and Uguru, J. O. (2011). "The impact of tone language and non-native language listening on measuring speech quality." In: *Journal of the Audio Engineering Society* 59.9, pp. 647–655.

Eisler, H. (1966). "Measurement of perceived acoustic quality of sound-reproducing systems by means of factor analysis." In: *Journal of the Acoustical Society of America* 39.3, pp. 484–492.

Engelke, U., Darcy, D. P., Mulliken, G. H., Bosse, S., Martini, M. G., Arndt, S., Antons, J. N., Chan, K. Y., Ramzan, N., and Brunnström, K. (2017). "Psychophysiology-based QoE assessment: A survey." In: *IEEE Journal of Selected Topics in Signal Processing* 11.1, pp. 6–21.

Evans, M., Ferne, T., Watson, Z., Melchior, F., Brooks, M., Stenton, P., and Forrester, I. (2016). "Creating object-based experiences in the real world." In: *International Broadcasting Convention (IBC) 2016 Conference*.

Faye, P., Brémaud, D., Teillet, E., Courcoux, P., Giboreau, A., and Nicod, H. (2006). "An alternative to external preference mapping based on consumer perceptive mapping." In: *Food Quality and Preference* 17.7, pp. 604 –614.

Fiebig, A. (2015). "Influence of context effects on sound quality assessments." In: *Proceedings of EuroNoise 2015*, pp. 2555–2560.

Focal (2014). *Dimension white paper*. Online, available from: https://www.focal.com/en/home-audio/home-theater/dimension/soundbar-dimension. Accessed: 2017-09-29.

Francombe, J., Brookes, T., Mason, R., and Woodcock, J. (2016). "Determining and labelling the preference dimensions of spatial audio replay." In: *8th International Conference on Quality of Multimedia Experience (QoMEX)*.

Francombe, J., Mason, R., Jackson, P. J. B., Brookes, T., Franck, A., Hughes, R., Woodcock, J., Melchior, F., and Pike, C. (2017a). "Media device orchestration for immersive spatial audio reproduction." In: *Proceedings of Audio Mostly (AM), London, UK*.

Francombe, J., Brookes, T., and Mason, R. (2017). "Evaluation of spatial audio reproduction methods (part 1): Elicitation of perceptual differences." In: *Journal of the Audio Engineering Society* 65.3, pp. 198–211.

Francombe, J., Brookes, T., Mason, R., and Woodcock, J. (2017b). "Evaluation of spatial audio reproduction methods (part 2): Analysis of listener preference." In: *Journal of the Audio Engineering Society* 65.3, pp. 212–225.

Gabrielsson, A. (1979). "Dimension analyses of perceived sound quality of sound-reproducing systems." In: *Scandinavian Journal of Psychology* 20.3, pp. 159–169.

Geerts, D., De Moor, K., Ketykó, I., Jacobs, A., Bergh, J. Van den, Joseph, W., Martens, L., and De Marez, L. (2010). "Linking an integrated framework with appropriate methods for measuring QoE." In: *2nd International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 158–163.

Geier, M., Wierstorf, H., Ahrens, J., Wechsung, I., Raake, A., and Spors, S. (2010). "Perceptual evaluation of focused sources in wave field synthesis." In: *Audio Engineering Society Convention 128*.

Gerzon, M. A. (1973). "Periphony: With-height sound reproduction." In: *Journal of the Audio Engineering Society* 21, pp. 2–10.

Giacalone, D., Nitkiewicz, M., Moulin, S., Boðason, T., Lund Laugesen, J., and Bech, S. (2017). "Sensory profiling of high-end loudspeakers using rapid methods - Part 2: Projective mapping with expert and naïve assessors." In: *Audio Engineering Society Convention 142*.

Glasberg, B. R. and Moore, B. C. J. (2002). "A model of loudness applicable to time-varying sounds." In: *Journal of the Audio Engineering Society* 50.5, pp. 331–342.

Goldsmith, R. E. and Hofacker, C. F. (1991). "Measuring consumer innovativeness." In: *Journal of the Academy of Marketing Science* 19.3, pp. 209–221.

Guastavino, C. (2003). "Étude sémantique et acoustique de la perception des basses fréquences dans l'environnemet sonore urbain (Semantic and acoustic approaches to low frequency perception)." PhD thesis. Université Paris.

Guastavino, C. and Katz, B. F. G. (2004). "Perceptual evaluation of multi-dimensional spatial audio reproduction." In: *Journal of the Acoustical Society of America* 116.2, pp. 1105–1115.

Guastavino, C., Katz, B. F. G., Polack, J.-D., Levitin, D. J., and Dubois, D. (2005). "Ecological validity of soundscape reproduction." In: *Acta Acustica united with Acustica* 91.2, pp. 333–341.

Hacihabiboglu, H., Sena, E. D., Cvetkovic, Z., Johnston, J., and Smith III, J. O. (2017). "Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics." In: *IEEE Signal Processing Magazine* 34.3, pp. 36–54.

Herre, J., Hilpert, J., Kuntz, A., and Plogsties, J. (2015). "MPEG-H audio - The new standard for universal spatial/3D audio coding." In: *Journal of the Audio Engineering Society* 62.12, pp. 821–830.

Hertz, B. F. (1981). "100 years with stereo - The beginning." In: *Journal of the Audio Engineering Society* 29.5, pp. 368–370, 372.

Hooley, T. (2006). "Single box surround sound." In: *Acoustical Science and Technology* 27.6, pp. 354–360.

Horbach, U., Karamustafaoglu, A., Pellegrini, R., Mackensen, P., and Theile, G. (1999). "Design and applications of a data-based auralization system for surround sound." In: *Audio Engineering Society Convention 106*.

Huang, Y., Zhou, W., and Du, Y. (2014). "Research on the user behavior-based QoE evaluation method for HTTP mobile streaming." In: *Ninth International Conference on Broadband and Wireless Computing, Communication and Applications*, pp. 47–51.

ISO (2000). *Quality management systems - Fundamentals and vocabulary (ISO 9000)*. International Organization for Standardization.

ISO (2005). *Quality management systems - Fundamentals and vocabulary (ISO 9000)*. International Organization for Standardization.

ISO (2010). *Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems (ISO 9241 — 210)*. International Organization for Standardization.

ISO (2012). *Sensory analysis - General guidelines for the selection, training and monitoring of selected assessors and expert sensory assessors (ISO 8586)*. International Organization for Standardization.

ITU-R (2001). *BS.1387-1 Method for objective measurements of perceived audio quality*. International Telecommunication Union.

ITU-R (2012a). *BS.775-3 Multichannel stereophonic sound system with and without accompanying picture*. International Telecommunication Union.

ITU-R (2012b). *BT.500-13 Methodology for the subjective assessment of the quality of television pictures*. International Telecommunication Union.

ITU-R (2014). *BS.2300-0 Methods for assessor screening*. International Telecommunication Union.

ITU-R (2015a). *BS. 2159-7 Multichannel sound technology in home and broadcasting applications*. International Telecommunication Union.

ITU-R (2015b). *BS.1116-3 Methods for the subjective assessment of small impairments in audio systems*. International Telecommunication Union.

ITU-R (2015c). *BS.1534-3 Method for the subjective assessment of intermediate quality level of audio systems*. International Telecommunication Union.

ITU-R (2015d). *BS.1770-4 Algorithms to measure audio programme loudness and true-peak audio level*. International Telecommunication Union.

ITU-T (1996). *P.800 Methods for subjective determination of transmission quality*. International Telecommunication Union.

ITU-T (2001). *P.862 Method for objective measurements of perceived audio quality*. International Telecommunication Union.

ITU-T (2008a). *P.910 Subjective video quality assessment methods for multimedia applications*. International Telecommunication Union.

ITU-T (2008b). *Rec E.800 Definitions of terms related to quality of service*. International Telecommunication Union.

ITU-T (2008c). *Rec P.10, Vocabulary for performance and quality of service, amendment 2: New definitions for Inclusion in Recommendation ITU-T P.10/G.100*. International Telecommunication Union.

ITU-T (2011). *P.863 Perceptual objective listening quality assessment*. International Telecommunication Union.

Jackson, P. J. B., Dewhirst, M., Conetta, R., Zielinski, S., Rumsey, F., Meares, D., Bech, S., and George, S. (2008). "QESTRAL (part 3): System and metrics for spatial quality prediction." In: *Audio Engineering Society Convention 125*.

Jekosch, U. (2004). "Basic concepts and terms of "quality", reconsidered in the context of product-sound quality." In: *Acta Acustica united with Acustica* 90.6, pp. 999–1006.

Jekosch, U. (2005). *Voice and speech quality perception: Assessment and evaluation*. Berlin: Springer.

Jillings, N., Moffat, D., De Man, B., and Reiss, J. D. (2015). "Web Audio Evaluation Tool: A browser-based listening test environment." In: *12th Sound and Music Computing Conference*.

Johnson, R. B. and Onwuegbuzie, A. J. (2004). "Mixed methods research: A research paradigm whose time has come." In: *Educational Researcher* 33.7, pp. 14–26.

Jumisko-Pyykkö, S. (2011). "User-centered quality of experience and its evaluation methods for mobile television." PhD thesis. Tampere University of Technology.

Jumisko-Pyykkö, S. and Häkkinen, J. (2008). "Profiles of the evaluators: Impact of psychographic variables on the consumer-oriented quality assessment of mobile television." In: *Proceedings of IS&T/SPIE's International Symposium on Electronic Imaging: Science and Technology: Multimedia on Mobile Devices*.

Jumisko-Pyykkö, S., Häkkinen, J., and Nyman, G. (2007). "Experienced quality factors - Qualitative evaluation approach to audiovisual quality." In: *Proceedings of SPIE 6507 - The International Society for Optical Engineering*.

Jumisko-Pyykkö, S. and Utriainen, T. (2011). "A hybrid method for quality evaluation in the context of use for mobile (3D) television." In: *Multimedia Tools and Applications* 55.2, pp. 185–225.

Jumisko-Pyykkö, S. and Utriainen, T. (2010). "User-centered quality of experience of mobile 3DTV: How to evaluate quality in the context of use?" In: *Proc. SPIE 7542, Multimedia on Mobile Devices*.

Jumisko-Pyykkö, S. and Vainio, T. (2010). "Framing the context of use for mobile HCI." In: *International Journal of Mobile Human Computer Interaction* 2.4, pp. 1–28.

Kähäri, K., Åslund, T., and Olsson, J. (2011). "Preferred sound levels of portable music players and listening habits among adults: A field study." In: *Noise and Health* 13.50, pp. 9–15.

Karrer, K., Glaser, C., Clemens, C., and Bruder, C. (2009). "Technikaffinität erfassen der Fragebogen TA-EG [Assessing technical affinity - the questionnaire TA-EG]." In: *Der Mensch im Mittelpunkt technischer Systeme. 8. Berliner Werkstatt Mensch-Maschine-Systeme*. Ed. by Lichtenstein, A., Stößel, C., and Clemens, C. Düsseldorf: VDI Verlag GmbH, pp. 196–201.

Kean, J., Johnson, E., and Sheffield, E. (2015). *Study of audio loudness range for consumers in various listening modes and ambient noise levels*. Online: http://www.aes.org/technical/documentDownloads.cfm?docID=523. Accessed: 2017-09-29.

Kelly, G. (1955). *The psychology of personal constructs*. New York: Norton.

Kim, C. (2014). "Object-based spatial audio: Concept, advantages, and challenges." In: *3D Future Internet Media*. Ed. by Kondoz, A. and Dagiuklas, T. New York, NY: Springer New York, pp. 79–84.

Kim, S., Bakker, R., and Ikeda, M. (2016). "Timbre preferences of four listener groups and the influence of their cultural backgrounds." In: *Audio Engineering Society Convention 140*.

Kim, S., Lee, Y. W., and Pulkki, V. (2010). "New 10.2-channel vertical surround system (10.2-VSS); Comparison study of perceived audio quality in various multichannel sound systems with height loudspeakers." In: *Audio Engineering Society Convention 129*.

Kjeldsen, A. D. (1998). "The measurement of personal preferences by repertory grid technique." In: *Audio Engineering Society Convention 104*.

Koivuniemi, K. and Zacharov, N. (2001). "Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training." In: *Audio Engineering Society Convention 111*.

Kreibig, S. D. (2010). "Autonomic nervous system activity in emotion: A review." In: *Biological Psychology* 84.3, pp. 394–421.

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., and Kort, J. (2009). "Understanding, scoping and defining user experience: A survey approach." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. Boston, MA, USA: ACM, pp. 719–728.

Lawless, H. T. and Heymann, H. (2010). *Sensory evaluation of food - Principles and practices*. Springer.

Le Callet, P., Möller, S., and Perkis, A., eds. (2013). *Qualinet white paper on definitions of quality of experience, version 1.2*. Lausanne, Switzerland: European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003).

Leech, B. L. (2002). "Asking questions: Techniques for semistructured interviews." In: *PS: Political Science and Politics* 35.4, pp. 665–668.

Letowski, T. (1989). "Sound quality assessment: Concepts and criteria." In: *Audio Engineering Society Convention 87*.

Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S. (2014). "A spatial audio quality inventory (SAQI)." In: *Acta Acustica United with Acustica* 100.5, pp. 984–994.

Lindau, A. and Weinzierl, S. (2012). "Assessing the plausibility of virtual acoustic environments." In: *Acta Acustica united with Acustica* 98.5, pp. 804–810.

Lokki, T., Pätynen, J., Kuusinen, A., and Tervo, S. (2012). "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles." In: *Journal of the Acoustical Society of America* 132.5, pp. 3148–3161.

Lorho, G. (2005a). "Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating." In: *Audio Engineering Society Convention 118*.

Lorho, G. (2005b). "Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction." In: *Audio Engineering Society Convention 119*.

Mann, M., Churnside, A. W. P., Bonney, A., and Melchior, F. (2013). "Object-based audio applied to football broadcasts." In: *Proceedings of the 2013 ACM International Workshop on Immersive Media Experiences*. ImmersiveMe '13. Barcelona, Spain: ACM, pp. 13–16.

Martens, W. L. and Zacharov, N. (2000). "Multidimensional perceptual unfolding of spatially processed speech I: Deriving stimulus space using INDSCAL." In: *Audio Engineering Society Convention 109*.

Martin, G. and Bech, S. (2005). "Attribute identification and quantification in automotive audio - Part 1: Introduction to the descriptive analysis technique." In: *Audio Engineering Society Convention 118*.

Mason, A., Jillings, N., Ma, Z., Reiss, J. D., and Melchior, F. (2015). "Adaptive audio reproduction using personalized compression." In: *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology - Cinema, Television and the Internet*.

Mason, W. and Suri, S. (2012). "Conducting behavioral research on Amazon's Mechanical Turk." In: *Behavior Research Methods* 44.1, pp. 1–23.

Mattila, V.-V. (2001). "Descriptive analysis of speech quality in mobile communications: Descriptive language development and external preference mapping." In: *Audio Engineering Society Convention 111*.

McCarthy, J. and Wright, P. (2004). *Technology as experience*. The MIT Press.

McKeeg, A. and McGrath, D. S. (1997). "Using auralization techniques to render 5.1 surround to binaural and transaural playback." In: *Audio Engineering Society Convention 102*.

Melchior, F., Churnside, A., and Spors, S. (2012). "Emerging technology trends in spatial audio." In: *SMPTE Journal* 121.6, pp. 95–100.

Midgley, D. F. and Dowling, G. R. (1978). "Innovativeness: The concept and its measurement." In: *Journal of Consumer Research* 4.4, pp. 229–242.

Molesworth, B. R. C., Burgess, M., and Kwon, D. (2013). "The use of noise cancelling headphones to improve concurrent task performance in a noisy environment." In: *Applied Acoustics.* 74.1, pp. 110 –115.

Möller, S. (2010). *Quality engineering: Qualität kommunikationstechnischer Systeme*. London: Springer.

Möller, S. and Raake, A., eds. (2014). *Quality of experience: Advanced concepts, applications and methods*. Springer.

Möller, S. and Raake, A. (2014). "Quality of experience: Terminology, methods and applications." In: *Praxis der Informationsverarbeitung und Kommunikation* 37.4, pp. 255–263.

Moncel, T. d. (1881). "The telephone at the Paris Opera." In: *Scientific America*, pp. 422–423.

Moulin, S., Nicol, T., and Gros, L. (2012). "Spatial audio quality in regard to 3D video." In: *Acoustics 2012*.

Nakayama, T., Miura, T., Kosaka, O., Okamoto, M., and Shiga, T. (1971). "Subjective assessment of multichannel reproduction." In: *Journal of the Audio Engineering Society* 19.9, pp. 744–751.

Neitzel, R., Gershon, R. R. M., Zeltser, M., Canton, A., and Akram, M. (2009). "Noise levels associated with New York City's mass transit systems." In: *American Journal of Public Health* 99.8, p. 1393.

Nicol, R., Gros, L., Colomes, C., Noisternig, M., Warusfel, O., Bahu, H., Katz, B. F. G., and Simon, L. S. R. (2014). "A roadmap for assessing the quality of experience of 3D audio binaural rendering." In: *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, pp. 100–106.

Nixon, T., Bonney, A., and Melchior, F. (2015). "A reference listening room for 3D audio research." In: *3rd International Conference on Spatial Audio (ICSA)*. Graz, Austria.

Norman, G. (2010). "Likert scales, levels of measurement and the "laws" of statistics." In: *Advances in Health Sciences Education* 15.5, pp. 625–632.

OED Online (2017). *"Psychographics"*. https://en.oxforddictionaries.com/definition/psychographics, accessed: 2017-11-03. Oxford University Press.

Olive, S., Welti, T., and McMullin, E. (2014). "The influence of listeners' experience, age, and culture on headphone sound quality preferences." In: *Audio Engineering Society Convention 137*.

Parizet, E. (2002). "Paired comparison listening tests and circular error rates." In: *Acta Acustica united with Acustica* 88.4, pp. 594–598.

Parmentier, M. (2015). "Sound board: Object-based audio." In: *Journal of the Audio Engineering Society* 63.7/8, pp. 659–660.

Paul, S. (2009). "Binaural recording technology: A historical review and possible future developments." In: *Acta Acustica united with Acustica* 95.5, pp. 767–788.

Pedersen, T. H. (2009). *Applied listening test*. SenseCamp09 presentation.

Pedersen, T. H. and Fog, C. L. (1998). "Optimisation of perceived product quality." In: *EuroNoise 98*. Vol. 2, pp. 633–638.

Pedersen, T. H. and Zacharov, N. (2008). "How many psycho-acoustic attributes are needed?" In: *Journal of the Audio Engineering Society* 123.5, pp. 3163–3163.

Pedersen, T. H. and Zacharov, N. (2015). "The development of a sound wheel for reproduced sound." In: *Audio Engineering Society Convention 138*.

Peryam, D. R. and Girardot, N. F. (1952). "Advanced taste test method." In: *Food Engineering* 24.194, pp. 58–61.

Pike, C., Melchior, F., and Tew, T. (2014). "Assessing the plausibility of non-individualised dynamic binaural synthesis in a small room." In: *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*.

Preece, J., Rogers, Y., and Sharp, H. (2002). *Beyond interaction design: Beyond human-computer interaction*. New York, NY, USA: John Wiley & Sons, Inc.

Pulkki, V. (1997). "Virtual sound source positioning using vector base amplitude panning." In: *Journal of the Acoustical Society of America* 45.6, pp. 456–466.

Quintero M, R. and Raake, A. (2012). "Is taking into account the subjects degree of knowledge and expertise enough when rating quality?" In: *4th International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 194–199.

Raake, A. (2006). *Speech quality of VoIP: Assessment and prediction*. John Wiley & Sons.

Raake, A. and Blauert, J. (2013). "Comprehensive modeling of the formation process of sound-quality." In: *5th International Workshop on Quality of Multimedia Experience, (QoMEX)*, pp. 76–81.

Rainer, B., Waltl, M., Cheng, E., Shujau, M., Timmerer, C., Davis, S., Burnett, I., Ritz, C., and Hellwagner, H. (2012). "Investigating the impact of sensory effects on the Quality of Experience and emotional response in web videos." In: *4th International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 278–283.

Reis, T., Carriço, L., and Duarte, C. (2009). "Mobile interaction: Automatically adapting audio output to users and contexts on communication and media control scenarios." In: *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments: 5th International Conference, UAHCI 2009. Proceedings, Part II*. Ed. by Stephanidis, C. Springer Berlin Heidelberg, pp. 384–393.

Reiss, J. D. (2016). "A meta-analysis of high resolution audio perceptual evaluation." In: *Journal of the Audio Engineering Society* 64.6, pp. 364–379.

Reiter, U. and De Moor, K. (2012). "Content categorization based on implicit and explicit user feedback: Combining self-reports with EEG emotional state analysis." In: *4th International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 266–271.

Robitza, W., Garcia, M. N., and Raake, A. (2015). "At home in the lab: Assessing audiovisual quality of HTTP-based adaptive streaming with an immersive test paradigm." In: *7th International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 1–6.

Roto, V., Law, E., Vermeeren, A., and Hoonhout, J., eds. (2011). *User experience white paper: Bringing clarity to the concept of user experience*. Result from Dagstuhl Seminar on Demarcating User Experience.

Rumsey, F. (1998). "Subjective assessment of the spatial attributes of reproduced sound." In: *Proceedings of the AES 15th International Conference on Audio, Acoustics and Small Space*, pp. 122–135.

Rumsey, F. (2001). *Spatial audio*. Focal Press.

Rumsey, F. (2002). "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm." In: *Journal of the Audio Engineering Society* 50.9, pp. 651–666.

Rumsey, F. (2006). "Spatial audio and sensory evaluation techniques - Context, history and aims." In: *Spatial Audio & Sensory Evaluation Techniques*. Guildford, UK.

Rumsey, F., Zielinski, S., Kassier, R., and Bech, S. (2005a). "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality." In: *Journal of the Acoustical Society of America* 118.2, pp. 968–976.

Rumsey, F., Zielinski, S., Kassier, R., and Bech, S. (2005b). "Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences." In: *Journal of the Acoustical Society of America* 117.6, pp. 3832–3840.

Rumsey, F., Zielinski, S., Jackson, P., Dewhirst, M., Conetta, R., George, S., Bech, S., and Meares, D. (2008). "QESTRAL (part 1): Quality evaluation of spatial transmission and reproduction using an artificial listener." In: *Audio Engineering Society Convention 125* 1, pp. 349–356.

Rumsey, F. (2015). "Immersive audio, objects, and coding." In: *Journal of the Audio Engineering Society* 63.5, pp. 394–398.

Ryan, R. M. and Deci, E. L. (2000). "Intrinsic and extrinsic motivations: Classic definitions and new directions." In: *Contemporary Educational Psychology*, pp. 54–67.

Sabine, W. C. (1900). "Reverberation." In: *The American Architect*.

Sackl, A., Schatz, R., and Raake, A. (2017). "More than I ever wanted or just good enough? User expectations and subjective quality perception in the context of networked multimedia services." In: *Quality and User Experience* 2.1, p. 3.

Sackl, A., Masuch, K., Egger, S., and Schatz, R. (2012). "Wireless vs. wireline shootout: How user expectations influence quality of experience." In: *4th International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 148–149.

Satongar, D., Pike, C., Lam, Y. W., and Tew, A. I. (2015). "The influence of headphones on the localization of external loudspeaker sources." In: *Journal of the Audio Engineering Society* 63.10, pp. 799–810.

Schinkel-Bielefeld, N. (2017). "Audio quality evaluation in MUSHRA tests - Influences between loop setting and a listeners' ratings." In: *Audio Engineering Society Convention 142*.

Schinkel-Bielefeld, N., Lotze, N., and Nagel, F. (2013). "Audio quality evaluation by experienced and inexperienced listeners." In: *Proceedings of Meetings on Acoustics* 19.

Schinkel-Bielefeld, N., Jiandong, Z., Yili, Q., Leschanowsky, A. K., and Shanshan, F. (2017). "Is it harder to perceive coding artifact in foreign language items? - A study with mandarin chinese and german speaking listeners." In: *Audio Engineering Society Convention 142*.

Schoeffler, M. (2017). "Overall listening experience - A new approach to subjective evaluation of audio." PhD thesis. Friedrich-Alexander-Universität Erlangen-Nürnberg.

Schoeffler, M., Adami, A., and Herre, J. (2014). "The Influence of up- and down-mixes on the overall listening experience." In: *Audio Engineering Society Convention 137*.

Schoeffler, M., Conrad, S., and Herre, J. (2014). "The influence of the single/multi-channel-system on the overall listening experience." In: *Proc. of the AES 55th Conference on Spatial Audio*. Helsinki, Finland.

Schoeffler, M., Edler, B., and Herre, J. (2013). "How much does audio quality influence ratings of overall listening experience?" In: *Proceedings of the 10th Annual Symposium on Computer Music Multidisciplinary Research*.

Schoeffler, M. and Herre, J. (2013). "About the impact of audio quality on overall listening experience." In: *Proceedings of Sound and Music Computing Conference*. Stockholm, Sweden, pp. 48–53.

Schoeffler, M. and Herre, J. (2014). "About the different types of listeners for rating the overall listening experience." In: *Proceedings of Sound and Music Computing Conference 2014*. Athens, Greece.

Schoeffler, M. and Herre, J. (2016). "The relationship between basic audio quality and overall listening experience." In: *The Journal of the Acoustical Society of America* 140.3, pp. 2101–2112.

Schoeffler, M., Silzle, A., and Herre, J. (2017). "Evaluation of spatial/3D audio: Basic audio quality vs. quality of experience." In: *IEEE Journal of Selected Topics in Signal Processing* 11.1, pp. 75–88.

Schoeffler, M., Gernert, J. L., Neumayer, M., Westphal, S., and Herre, J. (2015a). "On the validity of virtual reality-based auditory experiments: A case study about ratings of the overall listening experience." In: *Virtual Reality*, pp. 1–20.

Schoeffler, M., Stöter, F.-R., Edler, B., and Herre, J. (2015b). "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R Recommendation BS.1534 (MUSHRA)." In: *1st Web Audio Conference*. Paris, France.

Scriven, F. (2005). "Two types of sensory panels or are there more?" In: *Journal of Sensory Studies* 20.6, pp. 526–538.

Shaw, M. L. G. and Gaines, B. R. (1989). "Comparing conceptual structures: Consensus, conflict, correspondence and contrast." In: *Knowledge Acquisition* 1.4, pp. 341–363.

Shirley, B. and Oldfield, R. (2015). "Clean audio for TV broadcast: An object-based approach for hearing-impaired viewers." In: *Journal of the Audio Engineering Society* 63.4, pp. 245–256.

Shirley, B., Meadows, M., Malak, F., Woodcock, J., and Tidball, A. (2017). "Personalized object-based audio for hearing impaired TV viewers." In: *Journal of the Audio Engineering Society* 65.4, pp. 293–303.

Silzle, A., George, S., Habets, E. A. P., and Bachmann, T. (2011). "Investigation on the quality of 3D sound reproduction." In: *International Conference on Spatial Audio (ICSA)*.

Silzle, A., Neugebauer, B., George, S., and Plogsties, J. (2009). "Binaural processing algorithms: Importance of clustering analysis for preference tests." In: *Audio Engineering Society Convention 126*.

Soulodre, G. A., Grusec, T., Lavoie, M., and Thibault, L. (1998). "Subjective evaluation of state-of-the-art two-channel audio codecs." In: *Journal of the Audio Engineering Society* 46.3, pp. 164–177.

Spors, S., Wierstorf, H., Raake, A., Melchior, F., Frank, M., and Zotter, F. (2013). "Spatial sound with loudspeakers and its perception: A review of the current state." In: *Proceedings of the IEEE* 101.9, pp. 1920–1938.

Staelens, N., Broeck, W. Van den, Pitrey, Y., Vermeulen, B., and Demeester, P. (2012). "Lessons learned during real-life QoE assessment." In: *10th European Conference on Interactive TV and Video, Proceedings*. Berlin, Germany: Ghent University, Department of Information technology, pp. 1–4.

Staffeldt, H. (1974). "Correlation between subjective and objective data for quality loudspeakers." In: *Journal of the Audio Engineering Society* 22.6, pp. 402–419.

Stone, H., Sidel, J., Oliver, S., Woolsey, A., and Singleton, C. (1974). "Sensory evaluation by quantitative descriptive analysis." In: *Food Technology* 28.1.

Strohmeier, D. (2011). "Open profiling of quality: A mixed methods approach for audiovisual quality evaluations." PhD thesis. Technische Universität Ilmenau.

Strohmeier, D. (2012). "Open profiling of quality: A mixed methods research approach for audiovisual quality evaluations." In: *SIGMultimedia Rec.* 4.4, pp. 5–6.

Strohmeier, D., Jumisko-Pyykkö, S., and Eulenberg, K. (2011). "Open profiling of quality: Probing the method in the context of use." In: *3rd International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 7–12.

Strohmeier, D., Jumisko-Pyykkö, S., and Kunze, K. (2010). "Open profiling of quality: A mixed method approach to understanding multimodal quality perception." In: *Advances in MultiMedia* 2010, 3:1–3:17.

Strohmeier, D., Jumisko-Pyykkö, S., and Reiter, U. (2010). "Profiling experienced quality factors of audiovisual 3D perception." In: *2nd International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 70–75.

Strohmeier, D., Jumisko-Pyykkö, S., Kunze, K., and H., B. M. (2011). "The extended-OPQ method for user-centered quality of experience evaluation: A study for mobile 3D video broadcasting over DVB-H." In: *EURASIP Journal on Image and Video Processing* 2011, pp. 1–24.

Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., and Colomes, C. (2000). "PEAQ - The ITU standard for objective measurement of perceived audio quality." In: *Journal of the Audio Engineering Society* 48.1/2, pp. 3–29.

Toole, F. E. (1985). "Subjective measurements of loudspeaker sound quality and listener performance." In: *Journal of the Audio Engineering Society* 33.1-2, pp. 2–32.

Torcoli, M., Herre, J., Paulus, J., Uhle, C., Fuchs, H., and Hellmuth, O. (2017). "The adjustment / satisfaction test (A/ST) for the subjective evaluation of dialogue enhancement." In: *Audio Engineering Society Convention 143*.

Turnbull, R., Hughes, P., and Hoare, S. (2008). "Audio enhancement for portable device based speech applications." In: *Audio Engineering Society Convention 124*.

Walton, T., Evans, M., Kirk, D., and Melchior, F. (2016). "Does environmental noise influence preference of background-foreground audio balance?" In: *Audio Engineering Society Convention 141*.

Wechsung, I., Schulz, M., Engelbrecht, K.-P., Niemann, J., and Möller, S. (2011). "All users are (not) equal - The influence of user characteristics on perceived quality, modality choice and performance." In: *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*. Springer New York, pp. 175–186.

Werner, S. and Klein, F. (2014). "Influence of context dependent quality parameters on the perception of externalization and direction of an auditory event." In: *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*.

Widén, S., Båsjö, S., Möller, C., and Kähäri, K. (2017). "Headphone listening habits and hearing thresholds in Swedish adolescents." In: *Noise and Health* 19.88.

Wiggins, B. (2010). *WigWare*. Online: http://www.brucewiggins.co.uk/?page_id=78. Accessed: 2017-09-29.

Williams, A. A. and Langron, S. P. (1984). "The use of free-choice profiling for the evaluation of commercial ports." In: *Journal of the Science of Food and Agriculture* 35.5, pp. 558–568.

Woodcock, J., Moorhouse, A. T., and Waddington, D. C. (2014). "A multidimensional evaluation of the perception and annoyance caused by railway induced groundborne vibration." In: *Acta Acustica united with Acustica* 100.4, pp. 614–627.

Woodcock, J., Davies, W. J., Cox, T. J., and Melchior, F. (2016). "Categorization of broadcast audio objects in complex auditory scenes." In: *Journal of the Audio Engineering Society* 64.6, pp. 380–394.

Wright, P. and McCarthy, J. (2010). *Experience-centered design: Designers, users, and communities in dialogue*. Morgan and Claypool Publishers.

Yamaha (2012). *Yamaha YSP-4300 owner's manual*. Online, available from: http://download.yamaha.com/api/asset/file/?language=en&site=ca.yamaha.com&asset_id=59934. Accessed: 2017-09-29.

Zacharov, N. and Koivuniemi, K. (2001). "Unraveling the perception of spatial sound reproduction: Techniques and experimental design." In: *Audio Engineering Society Conference: 19th International Conference: Surround Sound - Techniques, Technology, and Perception*.

Zacharov, N., Pike, C., Melchior, F., and Worch, T. (2016). "Next generation audio system assessment using the multiple stimulus ideal profile method." In: *8th International Conference on Quality of Multimedia Experience (QoMEX)*.

Zielinski, S. K., Rumsey, F., and Bech, S. (2003). "Effects of bandwidth limitation on audio quality in consumer multichannel audiovisual delivery systems." In: *Journal of the Audio Engineering Society* 51.6, pp. 475–501.

Zielinski, S. K., Rumsey, F., Kassier, R., and Bech, S. (2005). "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." In: *Journal of the Audio Engineering Society* 53.3, pp. 174–192.

Zielinski, S., Rumsey, F., and Bech, S. (2002). "Subjective audio quality trade-offs in consumer multichannel audio-visual delivery systems. Part I: Effects of high frequency limitation." In: *Audio Engineering Society Convention 112*.

Zielinski, S., Rumsey, F., and Bech, S. (2008). "On some biases encountered in modern audio quality listening tests - A review." In: *Journal of the Audio Engineering Society* 56.6, pp. 427–451.

Zion Market Research (2016). *Soundbar market by type (2, 2.1, 5.1 and other) for music players, TV sets, computer systems and other applications: Global industry perspective, comprehensive analysis, size, share, growth, segment, trends and forecast, 2015 - 2021*. Online: https://www.zionmarketresearch.com/report/soundbar-market. Accessed: 2017-09-29.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` (modified) developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX.

Final Version as of May 19, 2018 (`classicthesis` version 2.0).